Geometry & Topology

msp

GEOMETRY & TOPOLOGY

msp.org/gt

See inside back cover or msp.org/gt for submission instructions.

On the top-weight rational cohomology of $\mathcal{A}_g$

Madeline Brandt

Juliette Bruce

Melody Chan

Margarida Melo

Gwyneth Moreland

Corey Wolfe

We compute the top-weight rational cohomology of $\mathcal{A}_g$ for $g = 5$, 6 and 7, and we give some vanishing results for the top-weight rational cohomology of $\mathcal{A}_8$, $\mathcal{A}_9$ and $\mathcal{A}_{10}$. When $g = 5$ and $g = 7$, we exhibit nonzero cohomology groups of $\mathcal{A}_g$ in odd degree, thus answering a question highlighted by Grushevsky. Our methods develop the relationship between the top-weight cohomology of $\mathcal{A}_g$ and the homology of the link of the moduli space of principally polarized tropical abelian varieties of rank g. To compute the latter we use the Voronoi complexes used by Elbaz-Vincent, Gangl and Soulé. In this way, our results make a precise connection between the rational cohomology of $\mathrm{Sp}_{2g}(\mathbb{Z})$ and $\mathrm{GL}_g(\mathbb{Z})$. Our computations also give natural candidates for compactly supported cohomology classes of $\mathcal{A}_g$ in weight 0 that produce the stable cohomology classes of the Satake compactification of $\mathcal{A}_g$ in weight 0, under the Gysin spectral sequence for the latter space.

1 Introduction

Let $\mathcal{A}_g$ be the moduli stack of principally polarized complex abelian varieties of dimension g. It is well known that $\mathcal{A}_g$ is a separated Deligne–Mumford stack, isomorphic to the quotient of the Siegel upper half-plane $\mathbb{H}_g$ under the action of the integral symplectic group $\mathrm{Sp}_{2g}(\mathbb{Z})$. Therefore $\mathcal{A}_g$ is smooth of dimension $d = \binom{g+1}{2}$, but it is not proper for $g > 0$. Since $\mathcal{A}_g$ is a complex algebraic variety, the rational cohomology groups of $\mathcal{A}_g$ admit a weight filtration in the sense of mixed Hodge theory, with graded pieces $\mathrm{Gr}_j^W H^\bullet(\mathcal{A}_g; \mathbb{Q})$ which may appear for j from 0 to $2d$. We refer to the piece of weight $2d$ as the *top-weight rational cohomology* of $\mathcal{A}_g$.

The orbifold Euler characteristic and the stable cohomology of $\mathcal{A}_g$ are classically understood; see for instance Borel [6] and Harder [29]. However, the full cohomology ring $H^\bullet(\mathcal{A}_g; \mathbb{Q})$ is a mystery even for small g. The cases when $g \leq 2$ are classically known, and the case when $g = 3$ is the work of Hain [27]. The full cohomology ring for $g \geq 4$ is already unknown, though when $g = 4$, much information can be

determined from Hulek and Tommasi [31; 32], where the complete Betti tables for both the Voronoi and the perfect cone compactifications of $\mathcal{A}_4$ are computed. In particular, the top-weight cohomology of $\mathcal{A}_4$ vanishes; see Remark 6.7.

We compute the top-weight rational cohomology of $\mathcal{A}_g$ for $2 \leq g \leq 7$. For $g = 3$ and 4, our computations agree with the above-mentioned results of Hain and of Hulek and Tommasi, respectively. Our first main result is then the following.

Theorem A   *The top-weight rational cohomology of $\mathcal{A}_g$ for $g = 5, 6$ and 7 is*

$$\mathrm{Gr}_{30}^W H^k(\mathcal{A}_5; \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } k = 15, 20, \\ 0 & \text{else}, \end{cases}$$

$$\mathrm{Gr}_{42}^W H^k(\mathcal{A}_6; \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } k = 30, \\ 0 & \text{else}, \end{cases}$$

$$\mathrm{Gr}_{56}^W H^k(\mathcal{A}_7; \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } k = 28, 33, 37, 42, \\ 0 & \text{else}. \end{cases}$$

This answers an open question of Grushevsky [24, Open Problem 7], who asked whether $\mathcal{A}_g$ ever has nonzero odd cohomology.

For broader context, recall that from the description of $\mathcal{A}_g$ as the quotient $[\mathbb{H}_g / \mathrm{Sp}_{2g}(\mathbb{Z})]$, it is a rational classifying space for the integral symplectic group $\mathrm{Sp}_{2g}(\mathbb{Z})$. Thus, $H^*(\mathcal{A}_g; \mathbb{Q}) \cong H^*(\mathrm{Sp}_{2g}(\mathbb{Z}); \mathbb{Q})$. The situation is analogous to that of the moduli space of curves $\mathcal{M}_g$, which is a rational classifying space for the mapping class group Mod_g via its action on Teichmüller space. Moreover, in both cases, we find ourselves in the advantageous situation that $\mathcal{M}_g$ and $\mathcal{A}_g$ are smooth and separated Deligne–Mumford stacks with coarse moduli spaces which are algebraic varieties, permitting Deligne's mixed Hodge theory to be applied to study the rational cohomology of these groups. The results of this paper use this algebrogeometric perspective to find new nonzero classes in a canonical quotient of $H^*(\mathrm{Sp}_{2g}(\mathbb{Z}); \mathbb{Q})$: the *top-weight* quotient, in the sense of mixed Hodge theory. (Recall that in general, the rational cohomology of a complex algebraic variety X of dimension d admits a weight filtration with graded pieces $\mathrm{Gr}_j^W H^k(X; \mathbb{Q})$. $\mathrm{Gr}_j^W H^k(X; \mathbb{Q})$ vanishes whenever $j > 2d$, so $\mathrm{Gr}_{2d}^W H^k(X; \mathbb{Q})$ is referred to as the *top-weight* part of $H^k(X; \mathbb{Q})$.)

Indeed, in this paper, we develop methods for studying $\mathcal{A}_g$ that are analogous to those employed in Chan, Galatius and Payne [12] for $\mathcal{M}_g$. The moduli spaces $\mathcal{A}_g$ admit toroidal compactifications $\overline{\mathcal{A}}_g^\Sigma$, which are proper Deligne–Mumford stacks; see Faltings and Chai [23, Theorem 5.7]. The compactifications $\overline{\mathcal{A}}_g^\Sigma$ are associated to admissible decompositions Σ of Ω_g^{rt}, the rational closure of the cone of positive definite quadratic forms in g variables; see Section 2.3. The same data is also used to construct the moduli space $A_g^{\mathrm{trop}, \Sigma}$ of tropical abelian varieties of dimension g in the category of generalized cone complexes; see Section 2.6.

Then for any admissible decomposition Σ of Ω_g^{rt} and for each $i \geq 0$, and writing $LA_g^{\text{trop},\Sigma}$ for the link of the cone point of $A_g^{\text{trop},\Sigma}$, there is a canonical identification

$$\tilde{H}_{i-1}(LA_g^{\text{trop},\Sigma};\mathbb{Q}) \cong \text{Gr}_{2d}^W H^{2d-i}(\mathcal{A}_g;\mathbb{Q}).$$

This statement can be deduced from Odaka and Oshima [40, Corollary 2.9] (see pages 24–25 of op. cit.); since the language is different, and in order to be self-contained, we give a short proof in Theorem 3.1. Briefly, there exist admissible decompositions Σ for which $\overline{\mathcal{A}}_g^\Sigma$ is a smooth simple normal crossings compactification of $\mathcal{A}_g$ whose boundary complex is identified with $LA_g^{\text{trop},\Sigma}$. However, the homeomorphism type of $LA_g^{\text{trop},\Sigma}$ is independent of Σ: see Section 3 or [39, Remark A.14]. The conclusion follows by applying the generalization to Deligne–Mumford stacks, spelled out in Chan, Galatius and Payne [12], of Deligne's comparison theorems in mixed Hodge theory; see Theorem 3.1.

We then compute the topology of $A_g^{\text{trop},\Sigma}$ by considering the *perfect* or *first Voronoi* toroidal compactification $\overline{\mathcal{A}}_g^P$ and its tropical version $A_g^{\text{trop},P}$, associated to the *perfect cone decomposition* (Fact 2.6). This decomposition is very well studied and enjoys interesting combinatorial properties, which are well suited for our computations. We identify the homology of the link of $A_g^{\text{trop},P}$ with the homology of the *perfect chain complex* $P_\bullet^{(g)}$ (Definition 4.1, Proposition 4.4), using the framework of cellular chain complexes of symmetric CW–complexes due to Allcock, Corey and Payne [3].

To compute the homology of the complex $P_\bullet^{(g)}$ we use a related complex $V_\bullet^{(g)}$, called the Voronoi complex. This was introduced in Elbaz-Vincent, Gangl and Soulé [22] and Lee and Szczarba [36] to compute the cohomology of the modular groups $\text{GL}_g(\mathbb{Z})$ and $\text{SL}_g(\mathbb{Z})$. They use the perfect form cell decomposition of Ω_g^{rt}, which is invariant under the action of each of these groups, and then relate the equivariant homology of Ω_g^{rt} modulo its boundary with the cohomology of $\text{GL}_g(\mathbb{Z})$ and $\text{SL}_g(\mathbb{Z})$, respectively. For this purpose, the homology of $V_\bullet^{(4)}$ was computed by Lee and Szczarba [36] for $\text{SL}_4(\mathbb{Z})$ (and we adapt this computation to the case of $\text{GL}_4(\mathbb{Z})$ in this paper), while for $g = 5, 6$ and 7 the complex $V_\bullet^{(g)}$ was computed in [22] with the help of a computer program using lists of perfect forms for $g \leq 7$ by Jaquet [34]. In Theorem 4.13 we show that the complexes $P_\bullet^{(g)}$ and $V_\bullet^{(g)}$ sit in an exact sequence

$$(1) \qquad\qquad 0 \to P_\bullet^{(g-1)} \to P_\bullet^{(g)} \xrightarrow{\pi} V_\bullet^{(g)} \to 0.$$

This sequence together with the results in [22] are then crucial to get our main result.

In Section 5 we consider a subcomplex of $P_\bullet^{(g)}$ called the inflation complex and prove that it is acyclic. Using this result, we show that $\text{Gr}_{g^2+g}^W H^i(\mathcal{A}_g;\mathbb{Q}) = 0$ for $i > g^2$, which recovers the vanishing in top weight of the rational cohomology of $\mathcal{A}_g$ in degree above the virtual cohomological dimension (which for $\mathcal{A}_g$ is equal to g^2).

For $g = 8, 9$ and 10, full calculations of the top-weight cohomology of $\mathcal{A}_g$ are beyond the scope of our computations. However, our computations for $g = 7$ together with a vanishing result of Dutour Sikirić, Elbaz-Vincent, Kupers and Martinet [21] allow us to deduce, in Section 6.5, the vanishing of $\text{Gr}_{(g+1)g}^W H^\bullet(\mathcal{A}_g;\mathbb{Q})$ in a range slightly larger than what is implied by virtual cohomological dimension bounds.

Theorem B *The top-weight rational cohomology of $\mathcal{A}_8$, $\mathcal{A}_9$ and $\mathcal{A}_{10}$ vanish in the following ranges*:

$$\mathrm{Gr}_{72}^W H^i(\mathcal{A}_8; \mathbb{Q}) = 0 \quad \text{for } i \geq 60,$$
$$\mathrm{Gr}_{90}^W H^i(\mathcal{A}_9; \mathbb{Q}) = 0 \quad \text{for } i \geq 79,$$
$$\mathrm{Gr}_{110}^W H^i(\mathcal{A}_{10}; \mathbb{Q}) = 0 \quad \text{for } i \geq 99.$$

To provide some broader context for our main results on $H^*(\mathcal{A}_g; \mathbb{Q}) \cong H^*(\mathrm{Sp}_{2g}(\mathbb{Z}); \mathbb{Q})$, we now highlight two interesting connections: first, to the stable cohomology of Satake compactifications, and second, to the cohomology of general linear groups $\mathrm{GL}_g(\mathbb{Z})$. More details appear in Section 7.

Relationship with the stable cohomology of $\mathcal{A}_g^{\mathrm{Sat}}$ By Poincaré duality, the top-weight cohomology of $\mathcal{A}_g$ studied in this paper admits a perfect pairing with weight 0 compactly supported cohomology of $\mathcal{A}_g$. These weight 0 classes, in turn, have an interesting, not yet fully understood relationship with the stable cohomology ring of the Satake compactification $\mathcal{A}_g^{\mathrm{Sat}}$, whose structure was first understood by Charney and Lee [14].

Indeed, the stable cohomology ring of $\mathcal{A}_g^{\mathrm{Sat}}$ is freely generated by extensions of the well-known odd λ–classes and by less-understood classes y_6, y_{10}, $y_{14}, \ldots$ which were proven to be of weight 0 by Chen and Looijenga in [15]. This predicts the existence of infinitely many top-weight cohomology classes of $\mathcal{A}_g$ as g grows. More precisely, the classes found in the present paper, with Poincaré duality applied, give natural candidates for the "sources" of the y_j in the sense of persisting in a Gysin spectral sequence relating the compactly supported cohomology groups of the space $\mathcal{A}_g^{\mathrm{Sat}}$ and those of the spaces $\mathcal{A}_{g'}$ for $g' \leq g$. See Table 4 at the end of the paper for a summary of everything that is known on the E_1 page of this spectral sequence in weight 0.

This connection was explained to us by O Tommasi and provides significant additional interest in our main results; we discuss it in detail in Section 7.

Relationship with the cohomology of $\mathrm{GL}_g(\mathbb{Z})$ Second, we would like to emphasize the connection between $H^*(\mathrm{Sp}_{2g}(\mathbb{Z}); \mathbb{Q})$ and $H^*(\mathrm{GL}_g(\mathbb{Z}); \widetilde{\mathbb{Q}})$ provided by our main results, where $\widetilde{\mathbb{Q}}$ denotes the orientation module on the link of the positive definite cone. The possibility of such a connection is essentially present in [4], but the precise connection employed in this paper, which is a key step in proving our main Theorems A and B, has been underutilized in the literature.

Indeed, Theorem 4.13 of this paper shows the exactness of the sequence (1) relating the perfect complexes $P^{(g-1)}$ and $P^{(g)}$ on the one hand, and the Voronoi complexes $V^{(g)}$. Again, these complexes are related to $H^*(\mathrm{Sp}_{2g}(\mathbb{Z}); \mathbb{Q})$ and $H^*(\mathrm{GL}_g(\mathbb{Z}); \mathbb{Q})$, respectively: precisely, for all k,

$$H^{\binom{g}{2}-k}(\mathrm{GL}_g(\mathbb{Z}); \widetilde{\mathbb{Q}}) \cong H_{k+g-1}(V^{(g)})$$

and

$$H_{k-1}(P^{(g)}) \cong \mathrm{Gr}_{g^2+g}^W H^{g^2+g-k}(\mathcal{A}_g; \mathbb{Q}) \twoheadleftarrow H^{g^2+g-k}(\mathcal{A}_g; \mathbb{Q}).$$

(See Soulé [43], Elbaz-Vincent, Gangl and Soulé [22, Section 3.4] and Proposition 4.4, respectively). For example, in view of exactness of (1), it is immediately possible to pass vanishing results on the top-weight quotient of $H^*(\mathcal{A}_g; \mathbb{Q})$ and vanishing results on $H^*(\mathrm{GL}_g(\mathbb{Z}); \widetilde{\mathbb{Q}})$ back and forth. For instance, recall that Church, Farb and Putman conjectured [16, Conjecture 2] that

$$H^{\binom{g}{2}-i}(\mathrm{SL}_g(\mathbb{Z}); \mathbb{Q}) = 0 \quad \text{for all } i < g - 1,$$

which implies the analogous statement for $\mathrm{GL}_g(\mathbb{Z})$ with both $\mathbb{Q}$ and $\widetilde{\mathbb{Q}}$ coefficients [22, equation (7)]. The conjecture is true for $i = 0$ by Lee and Szczarba [35], for $i = 1$ by Church and Putman [17], and for $i = 2$ by the recent preprint of Brück, Miller, Patzt, Sroka and Wilson [10], which appeared after the original version of this paper. As corollaries of these results and the results of this paper, we thus have:

Corollary 1.1 *For all $g > 0$,*

$$\mathrm{Gr}^W_{g^2+g} H^{g^2-k}(\mathcal{A}_g; \mathbb{Q}) = 0 \quad \text{when } k \leq 2.$$

This agrees with the $g = 9$ and $g = 10$ vanishing results in Theorem B. More generally, the Church–Farb–Putman conjecture would imply that

$$\mathrm{Gr}^W_{g^2+g} H^{g^2-i}(\mathcal{A}_g; \mathbb{Q}) = 0 \quad \text{whenever } i < g - 1.$$

That is, it would imply vanishing of $E_1^{p,q}$ in the spectral sequence in Table 4 for all $q < p - 1$; see Section 7.

It would be very interesting to find connections to the cohomology of $\mathrm{GL}_g(\mathbb{Z})$ that go deeper in the weight filtration on $H^*(\mathcal{A}_g; \mathbb{Q})$.

Organization of the paper In Section 2, we give the necessary preliminaries. This includes a discussion of generalized cone complexes, their links, and their homology. We then discuss admissible decompositions of the rational closure of the set of positive definite quadratic forms, and focus in particular on the perfect cone decomposition. We also give a brief introduction to matroids and to perfect cones associated to matroids. Then, we give some background on the tropical moduli space $\mathcal{A}_g^{\mathrm{trop},\Sigma}$, and on the construction of toroidal compactifications of $\mathcal{A}_g$ out of admissible decompositions.

In Section 3, we prove Theorem 3.1, which relates the top-weight cohomology of $\mathcal{A}_g$ to the reduced rational homology of the link of $\mathcal{A}_g^{\mathrm{trop},\Sigma}$. In Section 4, we show that the perfect chain complex $P_\bullet^{(g)}$ computes the top-weight cohomology of $\mathcal{A}_g$ (Proposition 4.4). We also relate this chain complex to the Voronoi complex $V_\bullet^{(g)}$ (Theorem 4.13). In Section 5, we introduce the inflation subcomplex, which we show is acyclic in Theorem 5.15. We prove an analogous result for the coloop subcomplex $C_\bullet^{(g)}$ of the regular matroid complex $R_\bullet^{(g)}$, which may be useful for future results.

In Section 6, we put together the results obtained in Section 4 with the computations of Lee and Szczarba [36] for $g = 4$ and Elbaz-Vincent, Gangl and Soulé [22] in $g = 5, 6$ and 7 to describe the top-weight cohomology of $\mathcal{A}_g$ for $g = 4, 5, 6$ and 7 and to give the above-mentioned bound for the vanishing of the cohomology of $\mathcal{A}_g$ in top weight for $g = 8, 9$ and 10. This proves Theorems A and B.

In Section 7, we discuss the relationship with the stable cohomology of the Satake compactification, including some open questions which are partially addressed by our main results and which deserve further attention.

2 Preliminaries

In this section we give preliminaries and introduce notation.

2.1 Cones and generalized cone complexes

A *rational polyhedral cone* σ in $\mathbb{R}^g$ (or just a *cone*, for simplicity) is the nonnegative real span of a finite set of integer vectors $v_1, v_2, \ldots, v_n \in \mathbb{Z}^g$,

$$\sigma := \mathbb{R}_{\geq 0}\langle v_1, v_2, \ldots, v_n \rangle := \left\{ \sum_{i=1}^{n} \lambda_i v_i : \lambda_i \in \mathbb{R}_{\geq 0} \right\}.$$

We assume all cones $\sigma \subset \mathbb{R}^g$ are *strongly convex*, meaning σ contains no nonzero linear subspaces of $\mathbb{R}^g$. The *dimension* of σ is the dimension of its linear span. The cone σ is said to be *smooth* if it is possible to choose the generating vectors $v_1, \ldots, v_n$ so that they are a subset of a $\mathbb{Z}$–basis of $\mathbb{Z}^g$. Note that some sources refer to what we call smooth cones as *basic* cones. A d–dimensional cone σ is said to be *simplicial* if it is generated by d vectors which are linearly independent over $\mathbb{R}$. A *face* of σ is any nonempty subset of σ that minimizes a linear functional on $\mathbb{R}^g$. Faces of σ are themselves rational polyhedral cones. A *facet* is a face of codimension one.

Given cones $\sigma \in \mathbb{R}^g$ and $\sigma' \in \mathbb{R}^{g'}$, a *morphism* $\sigma \to \sigma'$ is a continuous map from σ to σ' obtained as the restriction of a linear map $\mathbb{R}^g \to \mathbb{R}^{g'}$ sending $\mathbb{Z}^g$ to $\mathbb{Z}^{g'}$. A *face morphism* is a morphism of cones $\sigma \to \sigma'$ sending σ isomorphically to a face of σ'. Notice that isomorphisms of cones are examples of face morphisms. Denote by Cones the category of cones with face morphisms.

The one-dimensional faces of σ are called the *extremal rays* of σ, and there are only finitely many of these. Given an extremal ray ρ of σ, the semigroup $\rho \cap \mathbb{Z}^g$ is generated by a unique element u_ρ called the *ray generator* of ρ. An automorphism of a strongly convex cone permutes its finitely many ray generators, and is uniquely determined by this permutation. So, $\mathrm{Aut}(\sigma)$ is finite.

A *generalized cone complex* (see [1]) is a topological space with a presentation as a colimit $X := \varinjlim_{i \in \mathcal{I}} \sigma_i$ of an arbitrary diagram of cones $\sigma : \mathcal{I} \to$ Cones, in which all morphisms of cones are face morphisms. A morphism $(X = \varinjlim_{i \in \mathcal{I}} \sigma_i) \to (X' = \varinjlim_{i \in \mathcal{I}} \sigma_i')$ is a continuous map $f : X \to X'$ such that for each cone σ_i in the presentation of X, there exists a cone σ_j' in the presentation of X' and a morphism of cones $f_i : \sigma_i \to \sigma_j'$ such that the following diagram commutes:

$$
\begin{array}{ccc}
\sigma_i & \xrightarrow{\;f_i\;} & \sigma_j' \\
\downarrow & & \downarrow \\
X & \xrightarrow{\;f\;} & X'
\end{array}
$$

We remark that the category of generalized cone complexes is equivalent to the one of stacky fans as defined in [13, Definition 2.1.7].

2.2 Links of generalized cone complexes

For any cone $\sigma \subset \mathbb{R}^g$, define the *link* of σ at the origin to be the topological space $L\sigma = (\sigma - \{0\})/\mathbb{R}_{>0}$, where the action of $\mathbb{R}_{>0}$ is by scalar multiplication. Thus $L\sigma$ is homeomorphic to a closed ball of dimension $\dim \sigma - 1$. A face morphism of cones $\sigma \to \sigma'$ induces a morphism of links $L\sigma \to L\sigma'$, making L a functor from Cones to topological spaces.

Let $X = \varinjlim_{i \in \mathcal{I}} \sigma_i$ be a generalized cone complex, where $\sigma : \mathcal{I} \to$ Cones is a diagram of cones. We define the *link* of X as the colimit

$$LX = \varinjlim(L \circ \sigma).$$

Thus LX is a topological space, equipped with a colimit presentation as above. In fact, LX is a *symmetric CW–complex*, by [3, Example 2.4]. The definition of symmetric CW–complex generalizes the *symmetric Δ–complexes* of [12]. Roughly, a symmetric CW–complex is like a CW–complex, except with closed n–balls replaced by quotients thereof by finite subgroups of the orthogonal group $O(n)$.

Let X be a finite generalized cone complex, meaning that the indexing category $\mathcal{I}$ is equivalent to one with a finite number of objects and morphisms. We now write down a chain complex isomorphic to the *cellular chain complex* of LX, in the sense of [3, Section 4] and [12, Section 3], whose homology is identified with the singular homology of LX.

For each $p \geq -1$, let $\mathrm{Cones}_p(X)$ denote the finite groupoid whose objects are all $(p+1)$–dimensional faces of σ_i for all $i \in \mathcal{I}$, with a morphism $\tau \to \tau'$ for each isomorphism of cones $\phi \colon \tau \xrightarrow{\cong} \tau'$ such that the following diagram commutes:

$$\tau \xrightarrow{\phi} \tau'$$
$$\searrow \swarrow$$
$$X$$

Let τ be a cone in $\mathrm{Cones}_p(X)$. We make use of three compatible notions of orientation found in the literature:

  (i)   an orientation of τ is an orientation of $L\tau$ (see [36]),

  (ii) an orientation of τ is an orientation of the suspension of $L\tau$ (see [3]), and

  (iii) an orientation of τ is an orientation of $\mathbb{R}\tau$, the $\mathbb{R}$–linear span of τ; ie it is a choice of ordered basis for $\mathbb{R}\tau$, up to a change of basis with positive determinant (see [22]).

For the first two definitions, it is clear that an orientation on τ induces an orientation on the faces of τ as well. For the third definition, given a facet τ' of τ, the induced orientation on τ' is any one such that the quantity $\epsilon(\tau', \tau)$, defined as follows, is 1.

Let $B = (v_1, \ldots, v_n, v)$, where $B' = (v_1, \ldots, v_n)$ is an orientation of τ' and v is a ray generator of a ray of τ not contained in τ'. Set $\epsilon(\tau', \tau)$ to be the sign of the orientation of B in the oriented vector space $\mathbb{R}\tau$. Note that this sign does not depend on the choice of v. These definitions are compatible, in that a choice of orientation under one definition yields a choice of orientation under the other two, and under this correspondence a cone morphism $\tau \to \sigma$ is orientation-preserving under one definition if it is orientation-preserving under all three. Say that $\tau \in \mathrm{Cones}_p(X)$ is *alternating* if all automorphisms $\tau \to \tau$ in $\mathrm{Cones}_p(X)$ are orientation-preserving on τ.

Choose a set Γ_p of representatives of isomorphism classes of alternating cones in $\mathrm{Cones}_p(X)$, and for each $\tau \in \Gamma_p$ fix an orientation ω_τ of τ. If ρ' is a facet of τ, then ω_τ induces an orientation of ρ', which we denote by $\omega_\tau|_{\rho'}$. Let $C_p(LX)$ be the $\mathbb{Q}$–vector space with basis Γ_p. We define a differential

$$\partial \colon C_p(LX) \to C_{p-1}(LX)$$

by extending linearly on $C_p(LX)$ the following definition: given $\tau \in \Gamma_p$ and $\rho \in \Gamma_{p-1}$, set

$$\partial(\tau)_\rho = \sum_{\rho'} \eta(\rho', \rho),$$

where ρ' ranges over the facets of τ that are isomorphic in $\mathrm{Cones}_{p-1}(X)$ to ρ, and where $\eta(\rho', \rho) = \pm 1$ according to whether an isomorphism $\phi \colon \rho' \to \rho$ in $\mathrm{Cones}_p(X)$ takes the orientation $\omega_\tau|_{\rho'}$ to ω_ρ or $-\omega_\rho$. Note that $\eta(\rho', \rho)$ is well defined, ie independent of choice of ϕ, precisely because ρ is alternating.

Let $C_\bullet(LX)$ denote the complex

$$\cdots \xrightarrow{\partial} C_p(LX) \xrightarrow{\partial} C_{p-1}(LX) \xrightarrow{\partial} \cdots \xrightarrow{\partial} C_{-1}(LX) \to 0.$$

The main proposition in this subsection is the following.

Proposition 2.1 *Let X be a finite generalized cone complex. Then $C_\bullet(LX)$ is a complex, ie $\partial^2 = 0$.*

(i) *If X is connected, we have, for each $p \geq 0$,*

$$H_p(C_\bullet(LX)) \cong \tilde{H}_p(LX; \mathbb{Q}).$$

(ii) *More generally, for each $p > 0$, we have canonical isomorphisms*

$$H_p(C_\bullet(LX)) \cong H_p(LX; \mathbb{Q}),$$

and for $p = 0$ we have

$$H_0(C_\bullet(LX)) \cong \ker(H_0(LX; \mathbb{Q}) \to \mathbb{Q}\Gamma_{-1}).$$

Proposition 2.1 follows from [3, Theorem 4.2], by tracing through their definition of the cellular chain complex of LX. We give a self-contained proof sketch below.

Proof sketch Write $LX^{(p)}$ for the p–skeleton of LX, ie the union of the images of $L\sigma$ in X, for σ ranging over cones of dimension at most $p+1$ in X. By a standard argument analogous to the proof of [30, Theorem 2.2.27], the complex

(2) $$\cdots \to H_p(LX^{(p)}, LX^{(p-1)}; \mathbb{Q}) \xrightarrow{\delta_p} H_{p-1}(LX^{(p-1)}, LX^{(p-2)}; \mathbb{Q}) \to \cdots$$

has homology canonically identified with the singular homology of LX. Moreover,

$$H_p(LX^{(p)}, LX^{(p-1)}; \mathbb{Q}) \cong \bigoplus_\tau H_p\big((L\tau)/\operatorname{Aut}(\tau), (\partial L\tau)/\operatorname{Aut}(\tau); \mathbb{Q}\big),$$

where τ ranges over a set of representatives of isomorphism classes in $\mathrm{Cones}_p(X)$. Here $\operatorname{Aut}(\tau) = \mathrm{Iso}_{\mathrm{Cones}_p(X)}(\tau, \tau)$ is the automorphism group of τ in $\mathrm{Cones}_p(X)$. Since $|\operatorname{Aut}(\tau)|$ is invertible in $\mathbb{Q}$, it follows that

$$H_p(LX^{(p)}, LX^{(p-1)}; \mathbb{Q}) \cong \bigoplus_\tau H_p(L\tau, \partial L\tau; \mathbb{Q})_{\operatorname{Aut}(\tau)},$$

and for each $\tau \in \mathrm{Cones}_p(X)$, we have

$$H_p(L\tau, \partial L\tau; \mathbb{Q})_{\mathrm{Aut}(\tau)} \cong \begin{cases} \mathbb{Q} & \text{if } \tau \text{ is alternating,} \\ 0 & \text{else,} \end{cases}$$

which identifies $H_p(LX^{(p)}, LX^{(p-1)}; \mathbb{Q})$ with $C_p(LX)$.   □

Remark 2.2 A statement analogous to Proposition 2.1 holds with $\mathbb{Q}$ replaced by a commutative ring R, if the order of $\mathrm{Aut}(\tau)$ is invertible in R for each τ.

Remark 2.3 See also the proofs in [36, Section 3], as well as [22, Section 3.3], which are written in the special cases of the *Voronoi complexes* for $\mathrm{SL}_g(\mathbb{Z})$ and $\mathrm{GL}_g(\mathbb{Z})$, but apply essentially verbatim to prove Proposition 2.1. The Voronoi complex of $\mathrm{GL}_g(\mathbb{Z})$ plays an important role in this paper.

2.3 Admissible decompositions

We now introduce admissible decompositions of the rational closure of the set of positive definite quadratic forms, which are used in the construction of toroidal compactifications of the moduli space of abelian varieties, as well as in the construction of the moduli space of tropical abelian varieties.

We denote by $\mathbb{R}^{\binom{g+1}{2}}$ the vector space of quadratic forms in $\mathbb{R}^g$, which we identify with $g \times g$ symmetric matrices with coefficients in $\mathbb{R}$. We denote by Ω_g the cone in $\mathbb{R}^{\binom{g+1}{2}}$ of positive definite quadratic forms. We define the rational closure of Ω_g to be the set Ω_g^{rt} of positive semidefinite quadratic forms whose kernel is defined over $\mathbb{Q}$. The group $\mathrm{GL}_g(\mathbb{Z})$ acts on the vector space $\mathbb{R}^{\binom{g+1}{2}}$ of quadratic forms by $h \cdot Q := hQh^t$, where $h \in \mathrm{GL}_g(\mathbb{Z})$ and h^t is its transpose. The cones Ω_g and Ω_g^{rt} are preserved by this action of $\mathrm{GL}_g(\mathbb{Z})$.

Remark 2.4 A positive semidefinite quadratic form Q in $\mathbb{R}^g$ belongs to Ω_g^{rt} if and only if there exists $h \in \mathrm{GL}_g(\mathbb{Z})$ such that

$$hQh^t = \begin{pmatrix} Q' & 0 \\ 0 & 0 \end{pmatrix}$$

for some positive definite quadratic form Q' in $\mathbb{R}^{g'}$ with $0 \le g' \le g$; see [38, Section 8].

The cones Ω_g and Ω_g^{rt} are not polyhedral cones. However, one can consider decompositions of these spaces into rational polyhedral cones, as in the following definition.

Definition 2.5 [38, Lemma 8.3], [23, Chapter IV.2] An *admissible decomposition* of Ω_g^{rt} is a collection $\Sigma = \{\sigma_\mu\}$ of rational polyhedral cones of Ω_g^{rt} such that

   (i)   if σ is a face of $\sigma_\mu \in \Sigma$ then $\sigma \in \Sigma$,

   (ii) the intersection of two cones σ_μ and σ_ν of Σ is a face of both cones,

   (iii) if $\sigma_\mu \in \Sigma$ and $h \in \mathrm{GL}_g(\mathbb{Z})$ then $h\sigma_\mu h^t \in \Sigma$,

(iv)   the set of $\mathrm{GL}_g(\mathbb{Z})$–orbits of cones is finite, and

(v) $\bigcup_{\sigma_\mu \in \Sigma} \sigma_\mu = \Omega_g^{\mathrm{rt}}$.

We say that two cones $\sigma_\mu, \sigma_\nu \in \Sigma$ are equivalent if they are in the same $\mathrm{GL}_g(\mathbb{Z})$–orbit.

There are three known families of admissible decompositions of Ω_g^{rt} described for all g: the perfect cone decomposition, the second Voronoi decomposition, and the central cone decomposition; see [38, Chapter 8]. In this paper, we work with the perfect cone decomposition, which we now describe.

2.4   The perfect cone decomposition

Given a positive definite quadratic form Q, consider the set of nonzero integral vectors where Q attains its minimum,

$$M(Q) := \{\xi \in \mathbb{Z}^g \setminus \{0\} : Q(\xi) \le Q(\zeta) \text{ for all } \zeta \in \mathbb{Z}^g \setminus \{0\}\}.$$

The elements of $M(Q)$ are called the *minimal vectors* of Q. Let $\sigma[Q]$ denote the rational polyhedral subcone of Ω_g^{rt} given by the nonnegative linear span of the rank-one forms $\xi \cdot \xi^t \in \Omega_g^{\mathrm{rt}}$ for elements ξ of $M(Q)$, ie

(3) $\sigma[Q] := \mathbb{R}_{\ge 0} \langle \xi \cdot \xi^t \rangle_{\xi \in M(Q)}.$

The *rank* of the cone $\sigma[Q]$ is defined to be the maximum rank of an element of $\sigma[Q]$; in fact the rank of $\sigma[Q]$ is exactly the dimension of the span of $M(Q)$; see Lemma 4.8.

Fact 2.6 [44]   *The set of cones*

$$\Sigma_g^{\mathrm{P}} := \{\sigma[Q] : Q \text{ is a positive definite form on } \mathbb{R}^g\}$$

*is an admissible decomposition of Ω_g^{rt}, known as the **perfect cone decomposition**.*

The quadratic forms Q such that $\sigma[Q]$ has maximal dimension $\binom{g+1}{2}$ are called *perfect*, hence the name of this admissible decomposition.

Example 2.7   Let us compute Σ_2^{P}. In this case, there is a unique perfect form up to $\mathrm{GL}_2(\mathbb{Z})$–equivalence, namely

$$Q = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

One can compute that $M(Q) = \{(\pm 1, 0), (0, \pm 1), (\pm 1, \mp 1)\}$. Thus, up to $\mathrm{GL}_2(\mathbb{Z})$–equivalence, there is a unique perfect cone $\sigma[Q]$ of maximal dimension 3, with ray generators

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

One may check that for $i \in \{0, 1, 2\}$, all i–dimensional faces of $\sigma[Q]$ are $\mathrm{GL}_2(\mathbb{Z})$–equivalent; hence there is a unique perfect cone of each dimension, up to the action of $\mathrm{GL}_2(\mathbb{Z})$.

Remark 2.8   The cones $\sigma[Q] \in \Sigma_g^{\mathrm{P}}$ need not be simplicial for $g \ge 4$; see [38, page 93].

2.5 Matroidal perfect cones

We now give a brief introduction to matroids and their associated orbits of perfect cones. Further background on matroids can be found in [41].

Definition 2.9 A *matroid* $M = (E, \mathcal{C})$ on a finite set E is a subset $\mathcal{C} \subset \mathcal{P}(E) \setminus \{\varnothing\}$, called the set of *circuits* of M, satisfying the following axioms:

(C1) No proper subset of a circuit is a circuit.

(C2) If $C_1, C_2 \in \mathcal{C}$ are distinct and $c \in C_1 \cap C_2$, then $(C_1 \cup C_2) - \{c\}$ contains a circuit.

A matroid $M = (E, \mathcal{C})$ is said to be *simple* if it has no circuits of length 1 or 2. A matroid $M = (E, \mathcal{C})$ is called *representable* over a field $\mathbb{F}$ if there is a matrix A over $\mathbb{F}$ such that E bijects to the columns of A with the circuits $\mathcal{C}$ of M indexing the minimal linearly dependent sets of columns of A. The matrix A is known as an $\mathbb{F}$–representation of M. An *automorphism* of a matroid is a bijection $\phi : E \to E$ such that for any subset $C \subset E$, C is a circuit of M if and only if $\phi(C)$ is a circuit of M.

Definition 2.10 A matroid is *regular* if and only if it is representable over every field.

A matroid M being regular is equivalent to M being representable over $\mathbb{R}$ by a totally unimodular matrix (ie a matrix such that every minor is either -1, 0 or 1). The *rank* of a regular matroid M is the smallest number r such that M is representable over $\mathbb{R}$ by a $r \times n$ totally unimodular matrix for some n; see [41, Lemma 2.2.21, page 85].

Definition 2.11 Let G be a graph. The *graphic matroid* $M(G)$ is the matroid with ground set $E(G)$ whose circuits are subsets of $E(G)$ forming a simple cycle of G.

Since graphic matroids are regular, they are representable over fields of any characteristic. This can be seen directly by constructing the following matrix representing $M(G)$. Fix an orientation of the edges of G. Let $A(G)$ be the $|V(G)| \times |E(G)|$ matrix with entries

$$A(G)_{ij} = \begin{cases} 0 & \text{if } v_i \notin e_j, \\ -1 & \text{if } v_i \text{ is the head of } e_j, \\ 1 & \text{if } v_i \text{ is the tail of } e_j. \end{cases}$$

This matrix represents the matroid $M(G)$ over any field.

Construction 2.12 Given a simple, regular matroid M of rank $\leq g$, choose a $g \times n$ totally unimodular matrix A that represents M over $\mathbb{R}$. Denoting the columns of A by $v_1, v_2, \ldots, v_n$, we let $\sigma_A(M) \subset \Omega_g^{\mathrm{rt}}$ be the rational polyhedral cone

$$\sigma_A(M) := \mathbb{R}_{\geq 0} \langle v_1 v_1^t, v_2 v_2^t, \ldots, v_n v_n^t \rangle.$$

By [37, Theorem 4.2.1], the cone $\sigma_A(M)$ is a perfect cone in Σ_g^{P}.

The cone $\sigma_A(M)$ is uniquely determined by M up to the action of $\mathrm{GL}_g(\mathbb{Z})$. In particular, if A and A' are two different totally unimodular matrices representing M over $\mathbb{R}$ then there exists an element $h \in \mathrm{GL}_g(\mathbb{Z})$ such that $h\sigma_A(M)h^t = \sigma_{A'}(M)$; see [37, Lemma 4.0.5(ii)]. We therefore denote the $\mathrm{GL}_g(\mathbb{Z})$–orbit of $\sigma_A(M)$ by $\sigma(M)$.

In the case of graphic matroids, Construction 2.12 can be made very explicit. As this is useful in Section 6, we take the time to explain it here. Fix $g > 0$. We now construct cones of Σ_g^{P} from graphs on $g + 1$ vertices. The rows of the $(g+1) \times |E(G)|$ matrix $A(G)$ as constructed above are linearly dependent. Let $A^*(G)$ be the matrix obtained from $A(G)$ by deleting the last row. The matrices $A(G)$ and $A^*(G)$ are both representations of $M(G)$. Let $v_1, \ldots, v_d$ be the columns of $A^*(G)$. Then $\sigma(M(G)) := \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_d v_d^t \rangle \in \Sigma_g^{\mathrm{P}}$ is a perfect cone; see [37, Theorem 4.2.1].

Definition 2.13 The *principal cone* is $\sigma_g^{\mathrm{prin}} := \sigma(M(K_{g+1}))$, the cone corresponding to the complete graph K_{g+1}.

When $g = 2$, this is the cone discussed in Example 2.7. More generally, for arbitrary g the principal cone can be defined as the cone corresponding to the quadratic form

$$\begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & 1 & \cdots & \frac{1}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & \frac{1}{2} & \cdots & 1 \end{bmatrix}.$$

These two definitions agree by [8, Lemma 6.1.3].

The faces of σ_g^{prin} may be understood as follows. Since $M(K_{g+1})$ is a simple matroid, the principal cone in Σ_g^{P} is simplicial by [8, Theorem 4.4.4(iii)]. Therefore, a codimension i face of the principal cone comes from a graph obtained by deleting i edges from K_{g+1}.

Remark 2.14 Automorphisms of the graph G give automorphisms of the matroid $M(G)$, but not all automorphisms of $M(G)$ arise in this way. However, if G is 3–connected, then $\mathrm{Aut}(G) = \mathrm{Aut}(M(G))$ (this is proved by Whitney in [45]; see [28, Lemmas 1 and 2]). The group $\mathrm{Aut}(M(G))$ is isomorphic to the group of permutations of the rays of $\sigma(M(G))$ induced by elements of $\mathrm{GL}_g(\mathbb{Z})$ stabilizing $\sigma(M(G))$ [11, Theorem 5.10].

2.6 The tropical moduli space A_g^{trop}

We now introduce the moduli space of tropical abelian varieties, which is a generalized cone complex constructed in [8] and later worked out in [13]. Our aim is to compute the homology of the link of A_g^{trop}, as this is canonically isomorphic to the top-weight rational cohomology of $\mathcal{A}_g$; see Theorem 3.1.

Definition 2.15 A *principally polarized tropical abelian variety* (or, for simplicity, just *tropical abelian variety*) of dimension g is a pair $A = (\mathbb{R}^g/\mathbb{Z}^g, Q)$, where Q is a positive semidefinite symmetric bilinear form on $\mathbb{R}^g$ with rational null space. We say that $A = (\mathbb{R}^g/\mathbb{Z}^g, Q)$ is *pure* if Q is positive definite.

Two tropical abelian varieties $(\mathbb{R}^g/\mathbb{Z}^g, Q)$ and $(\mathbb{R}^g/\mathbb{Z}^g, Q')$ are *isomorphic* if there is an $h \in \mathrm{GL}_g(\mathbb{Z})$ such that $Q' = hQh^t$. The set of isomorphism classes of tropical abelian varieties of dimension g is in bijective correspondence with the orbits in $\Omega_g^{\mathrm{rt}}/\mathrm{GL}_g(\mathbb{Z})$.

Given an admissible decomposition Σ of Ω_g^{rt}, define a generalized cone complex $A_g^{\mathrm{trop},\Sigma}$ by considering the stratified quotient of Ω_g^{rt} with respect to Σ; see [13, Definition 2.2.2]. Precisely, $A_g^{\mathrm{trop},\Sigma}$ is the generalized cone complex obtained as the colimit

$$A_g^{\mathrm{trop},\Sigma} := \varinjlim \{\sigma\}_{\sigma \in \Sigma}$$

with arrows given by inclusion of faces composed with the action of the group $\mathrm{GL}_g(\mathbb{Z})$ on Ω_g^{rt}: given two cones σ_i and $\sigma_j \in \Sigma$ and $h \in \mathrm{GL}_g(\mathbb{Z})$ with $h\sigma_i h^t$ a face of σ_j, we consider its associated lattice-preserving linear map $L_{i,j,g} \colon \sigma_i \hookrightarrow \sigma_j$ in the diagram. The space $A_g^{\mathrm{trop},\Sigma}$ is the *moduli space of tropical abelian varieties* of dimension g with respect to Σ.

2.7 Toroidal compactifications of the moduli space $\mathcal{A}_g$

In this paper, $\mathcal{A}_g$ denotes the moduli stack of principally polarized abelian varieties of dimension g. It is a smooth Deligne–Mumford algebraic stack of dimension $d = \binom{g+1}{2}$, and the coarse moduli space of principally polarized abelian varieties, denoted by A_g, is a quasiprojective variety.

The moduli stack $\mathcal{A}_g$ is not proper for $g > 0$, and there are different constructions of compactifications of $\mathcal{A}_g$. In particular, it is possible to construct normal crossings compactifications of $\mathcal{A}_g$ via the theory of toroidal compactifications. Both the constructions of $\mathcal{A}_g$ and of its toroidal compactifications as algebraic stacks were achieved in [4] over the complex numbers and in [23] over an arbitrary base. Even though we work over the complex numbers, we often refer to the constructions in [23] as these are more conveniently stated within the algebraic category and specifically for moduli of abelian varieties (rather than quotients of bounded symmetric domains as in [4]).

Let Σ be an admissible decomposition of Ω_g^{rt} (in the sense of Definition 2.5). Then one may associate to Σ a *toroidal compactification* $\overline{\mathcal{A}}_g^\Sigma$ of $\mathcal{A}_g$, which is a proper Deligne–Mumford stack, although in general it is not smooth. The fact that $\mathcal{A}_g \subset \overline{\mathcal{A}}_g^\Sigma$ is toroidal means that $(\mathcal{A}_g, \overline{\mathcal{A}}_g^\Sigma)$ is étale-locally isomorphic to a torus inside a toric variety.

By construction, the toroidal compactification $\overline{\mathcal{A}}_g^\Sigma$ comes with a stratification into locally closed subsets. These are in order-reversing bijection, with respect to the order relation given by the closure, with the $\mathrm{GL}_g(\mathbb{Z})$–equivalence classes of the relative interiors of the cones in Σ. For example, the origin of Ω_g^{rt}, which is the unique zero-dimensional cone in every admissible decomposition Σ, corresponds to the open

substack $\mathcal{A}_g$, which is the unique stratum of $\overline{\mathcal{A}}_g^\Sigma$ of maximal dimension d. At the other extreme, the maximal dimensional cones in Σ correspond to the zero-dimensional strata of $\overline{\mathcal{A}}_g^\Sigma$.

We study the *perfect* toroidal compactification $\mathcal{A}_g^{\text{perf}} = \overline{\mathcal{A}}_g^{\Sigma_g^P}$ of $\mathcal{A}_g$, ie the toroidal compactification of $\mathcal{A}_g$ associated to the perfect cone decomposition Σ_g^P. The geometric significance of the perfect cone compactification was highlighted in work of Shepherd-Barron [42], who shows that $\mathcal{A}_g^{\text{perf}}$ is the canonical model of $\mathcal{A}_g$ for $g \geq 12$. For our purposes it is particularly nice because the number of strata of codimension l in the boundary of $\mathcal{A}_g^{\text{perf}} \setminus \mathcal{A}_g$ is independent of g if $l \leq g$; see [26, Proposition 7.1].

3 A comparison theorem for $\mathcal{A}_g$ and A_g^{trop}

Let Σ be any admissible decomposition of Ω_g^{rt}. As mentioned above, we can use Σ to construct both a toroidal compactification of $\mathcal{A}_g$, and the generalized cone complex $A_g^{\text{trop},\Sigma}$, the moduli space of tropical abelian varieties associated to Σ. In this section, we record the relationship between the homology of $A_g^{\text{trop},\Sigma}$ with the top-weight cohomology of $\mathcal{A}_g$, as deduced from Deligne's comparison theorems and the framework in [12]. This precise relationship was already remarked by Odaka and Oshima [40, Corollary 2.9], as we explain further in Remark 3.2, but it is useful to have a self-contained proof, below.

Theorem 3.1 *For each $i \geq 0$ and admissible decomposition Σ, we have a canonical isomorphism*

$$\widetilde{H}_{i-1}(LA_g^{\text{trop},\Sigma}; \mathbb{Q}) \cong \text{Gr}_{2d}^W H^{2d-i}(\mathcal{A}_g; \mathbb{Q}),$$

where $d = \binom{g+1}{2}$ is the complex dimension of $\mathcal{A}_g$.

Proof First, by replacing Σ with another admissible decomposition of Ω_g^{rt} that refines it, we may assume that every cone of Σ is smooth and that it enjoys the following additional property: for any $h \in \text{GL}_g(\mathbb{Z})$ and $\sigma \in \Sigma$, we have that h fixes, pointwise, the cone $h\sigma h^t \cap \sigma$. Such a refinement is well known to exist [23, Chapter IV.2, page 98]. For example, one may be obtained by taking the barycentric refinement, which is simplicial, and then taking an appropriate *smooth* refinement which can be constructed as in [18, Theorem 11.1.9]. The homeomorphism type of $LA_g^{\text{trop},\Sigma}$ is unchanged when passing to a refinement.

Then, by [23, Theorem 5.7], it follows that $\overline{\mathcal{A}}_g^\Sigma$ is a smooth, separated Deligne–Mumford stack which is a simple normal crossings compactification of $\mathcal{A}_g$ and whose boundary complex is $LA_g^{\text{trop},\Sigma}$. Now the desired result follows from the following comparison theorem: we have a canonical isomorphism

$$\widetilde{H}_{i-1}(\Delta(\mathcal{X} \subset \overline{\mathcal{X}}); \mathbb{Q}) \cong \text{Gr}_{2d}^W H^{2d-i}(\mathcal{X}; \mathbb{Q})$$

for any normal crossings compactification $\mathcal{X} \subset \overline{\mathcal{X}}$ of smooth, separated Deligne–Mumford stacks over $\mathbb{C}$, where $\Delta(\mathcal{X} \subset \overline{\mathcal{X}})$ denotes the boundary complex of the pair $(\mathcal{X}, \overline{\mathcal{X}})$ and $d = \dim \mathcal{X}$ is the complex dimension of $\mathcal{X}$. This comparison theorem follows from Deligne's mixed Hodge theory [19; 20] in the case of complex varieties; we refer to [12] for the generalization to Deligne–Mumford stacks. $\square$

Remark 3.2 Let us briefly explain how Theorem 3.1 appears in Odaka [39] and Odaka and Oshima [40], since the language of those papers is somewhat different. Odaka and Oshima study certain "hybrid" compactifications of arithmetic quotients $\Gamma \backslash D$ of Hermitian symmetric domains. The case of $\mathcal{A}_g$ is the case $\Gamma = \mathrm{Sp}(2g, \mathbb{Z})$ and D is the "unit disc" of complex symmetric matrices Z with $Z^t \bar{Z} < \mathrm{Id}_g$. The point is that the boundary of these compactifications is homeomorphic to $L\mathcal{A}_g^{\mathrm{trop}, \Sigma}$, so the comparison statement in [40, Corollary 2.9], which relies on [12], combined with [12, Theorem 2.1], reduces to Theorem 3.1 in this case.

It is worth emphasizing the independence of choice of the admissible decomposition of Σ, as remarked in [39, A.14], that was implicit in the discussion above. More precisely, for any two admissible decompositions Σ_1 and Σ_2 of Ω_g^{rt}, we have a homeomorphism of links

$$L A_g^{\mathrm{trop}, \Sigma_1} \cong L A_g^{\mathrm{trop}, \Sigma_2}.$$

Indeed, it is well known that any two admissible decompositions Σ_1 and Σ_2 admit a common refinement $\widetilde{\Sigma}$ which is an admissible decomposition [23, Chapter IV.2, page 97], and by the construction of Section 2.2, we have canonical homeomorphisms

$$L A_g^{\mathrm{trop}, \Sigma_1} \cong L A_g^{\mathrm{trop}, \widetilde{\Sigma}} \cong L A_g^{\mathrm{trop}, \Sigma_2}.$$

4 The perfect and Voronoi chain complexes

In computing the top-weight cohomology of $\mathcal{A}_g$ there are two chain complexes that play central roles: the perfect chain complex $P_\bullet^{(g)}$ and the Voronoi chain complex $V_\bullet^{(g)}$. In this section we define both of these complexes, and show that the homology of the perfect chain complex $P_\bullet^{(g)}$ computes the top-weight cohomology of $\mathcal{A}_g$. Further, we show that the perfect and Voronoi complexes are related via a short exact sequence of chain complexes, which is useful as the Voronoi complex has seen more extensive study; see [22; 21]. We make use of this short exact sequence to prove our main results in Section 6.

4.1 The perfect chain complex

We first fix some notation, most of which we adapt from Section 2.2. For $n \in \mathbb{Z}$, let $\Sigma_g^{\mathrm{P}}[n]$ be the set of perfect cones in Σ_g^{P} of dimension $n + 1$, and denote the finite set of $\mathrm{GL}_g(\mathbb{Z})$–orbits of such cones by $\Sigma_g^{\mathrm{P}}[n]/\mathrm{GL}_g(\mathbb{Z})$. We write $\sigma \sim \sigma'$ if and only if σ and σ' lie in the same $\mathrm{GL}_g(\mathbb{Z})$–orbit. Recall that a cone $\sigma \in \Sigma_g^{\mathrm{P}}$ is *alternating* if and only if every element of $\mathrm{GL}_g(\mathbb{Z})$ stabilizing σ induces an orientation-preserving cone morphism of σ. If σ is an alternating cone then every cone in the same $\mathrm{GL}_g(\mathbb{Z})$–orbit as σ is alternating. We call such $\mathrm{GL}_g(\mathbb{Z})$–orbits alternating. Let $\Gamma_n^{(g)} = \Gamma_n$ be a set of representatives for the alternating elements of $\Sigma_g^{\mathrm{P}}[n]/\mathrm{GL}_g(\mathbb{Z})$.

For each n and each $\sigma \in \Gamma_n$, choose an orientation ω_σ on σ; the $\mathrm{GL}_g(\mathbb{Z})$–action extends this choice to a choice of orientation on every alternating cone in Σ_g^{P}. If $\rho \subset \sigma$ is an alternating facet of σ, denote the

orientation induced on ρ by $\omega_\sigma|_\rho$. Now let $\eta(\rho, \sigma)$ be 1 if the orientation on ρ agrees with the orientation induced by σ (ie $\omega_\rho = \omega_\sigma|_\rho$) and -1 otherwise. Finally, given $\sigma \in \Gamma_n$ and $\sigma' \in \Gamma_{n-1}$ define

$$(4) \qquad \delta(\sigma', \sigma) := \sum_{\rho \subset \sigma, \ \rho \sim \sigma'} \eta(\rho, \sigma),$$

where the sum is over all facets ρ of σ in the same $\mathrm{GL}_g(\mathbb{Z})$–orbit as σ'. With this notation in hand we can now define the perfect chain complex.

Definition 4.1 The *perfect chain complex* $(P_\bullet^{(g)}, \partial_\bullet)$ is the rational complex defined as follows. For each n, $P_n^{(g)}$ is the $\mathbb{Q}$–vector space with basis indexed by Γ_n. The differential $\partial_n \colon P_n^{(g)} \to P_{n-1}^{(g)}$ is given by

$$\partial(e_\sigma) := \sum_{\sigma' \in \Gamma_{n-1}} \delta(\sigma', \sigma) e_{\sigma'}.$$

Notice that $P_n^{(g)}$ is only possibly nonzero in the range $-1 \le n \le \binom{g+1}{2} - 1$, but even within this range $P_n^{(g)}$ may be zero, since alternating perfect cones do not necessarily exist in every dimension; see Example 4.2. While in many cases $\delta(\sigma', \sigma)$ is equal to $-1, 0$ or 1, this need not always be the case since a cone may have two or more facets that are $\mathrm{GL}_g(\mathbb{Z})$–equivalent.

Example 4.2 When $g = 2$, recall from Example 2.7 that up to the action of $\mathrm{GL}_2(\mathbb{Z})$ there is precisely one cone of maximal dimension,

$$\sigma_3 := \mathbb{R}_{\ge 0} \left\langle \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right\rangle.$$

Since σ_3 is simplicial, its faces correspond to all subsets of the above ray generators. One can show that up to the action of $\mathrm{GL}_2(\mathbb{Z})$ there is at most one cone in each dimension:

$$\sigma_2 := \mathbb{R}_{\ge 0} \left\langle \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\rangle, \quad \sigma_1 := \mathbb{R}_{\ge 0} \left\langle \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right\rangle, \quad \sigma_0 := \mathbb{R}_{\ge 0} \left\langle \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right\rangle.$$

Thus, to determine $\Gamma_{-1}, \Gamma_0, \Gamma_1$ and Γ_2, it is enough to see which of $\sigma_0, \sigma_1, \sigma_2$ and σ_3 are alternating. Consider the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

One can show that A stabilizes both σ_2 and σ_3 and that the induced cone morphism is orientation-reversing. Thus, Γ_1 and Γ_2 are empty. On the other hand, since the action of $\mathrm{GL}_2(\mathbb{Z})$ fixes the cone point σ_0, both σ_0 and σ_1 are alternating. So, $\Gamma_{-1} = \{\sigma_0\}$ and $\Gamma_0 = \{\sigma_1\}$. From this we see that the complex $P_\bullet^{(2)}$ is

$$
\begin{array}{ccccccccc}
P_2^{(2)} & & P_1^{(2)} & & P_0^{(2)} & & P_{-1}^{(2)} & & P_{-2}^{(2)} \\
& & & & & & & & \\
0 & \longrightarrow & 0 & \longrightarrow & \mathbb{Q}\langle e_{\sigma_1} \rangle & \xrightarrow{\ \partial_0\ } & \mathbb{Q}\langle e_{\sigma_0} \rangle & \longrightarrow & 0
\end{array}
$$

where ∂_0 sends e_{σ_1} to either e_{σ_0} or $-e_{\sigma_0}$ depending on the chosen orientations.
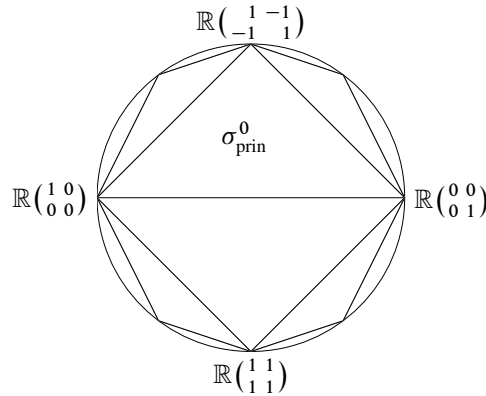
Figure 1: A section of Ω_2^{rt} and its perfect cone decomposition.

Remark 4.3 To be precise, the perfect complex $P_\bullet^{(g)}$ as constructed in Definition 4.1 is only unique up to isomorphism. In particular, the choice of representatives for Γ_n or reference orientations may result in different but isomorphic chain complexes. For instance, in Example 4.2, the differential ∂_0 is only determined up to sign.

The next proposition shows that the perfect complex $P_\bullet^{(g)}$ is isomorphic to the cellular chain complex associated to the symmetric CW complex $LA_g^{\mathrm{trop,P}}$. Thus, by Theorem 3.1, the homology of $P_\bullet^{(g)}$ computes the top-weight rational cohomology of $\mathcal{A}_g$.

Proposition 4.4 *For each $i \geq 0$, there exist canonical isomorphisms*

$$H_{i-1}(P_\bullet^{(g)}) \cong \widetilde{H}_{i-1}(LA_g^{\mathrm{trop,P}}; \mathbb{Q}) \cong \mathrm{Gr}_{2d}^W H^{2d-i}(\mathcal{A}_g; \mathbb{Q}),$$

where $A_g^{\mathrm{trop,P}} = A_g^{\mathrm{trop},\Sigma_g^P}$ is the tropical moduli space constructed in Section 2.6.

Proof By construction, $P_\bullet^{(g)}$ is naturally isomorphic to the cellular chain complex of $LA_g^{\mathrm{trop,P}}$ as defined in Section 2.2. Observe that the space $A_g^{\mathrm{trop,P}}$ is connected since it deformation retracts to the cone point. Thus, the first isomorphism then follows from part (i) of Proposition 2.1 and the second isomorphism follows from Theorem 3.1.   □

Example 4.5 By Example 4.2, we see that $P_\bullet^{(2)}$ has trivial homology in all degrees. Thus, Proposition 4.4 recovers the fact that $\mathcal{A}_2$ has trivial top-weight cohomology [33].

4.2 The Voronoi complex

Now we introduce a closely related complex, called the Voronoi complex $V_\bullet^{(g)}$, as considered in [22; 21].[1] We shall soon see that $V_\bullet^{(g)}$ is a quotient of $P_\bullet^{(g)}$, obtained by setting to zero the generators corresponding to cones contained in the boundary of Ω_g^{rt}.

---

[1]In [22; 21] the Voronoi complex is defined as a complex of free $\mathbb{Z}$–modules, while our definition of Voronoi complex is as a complex of $\mathbb{Q}$–vector spaces.

For each $n \in \mathbb{Z}$, let $\overline{\Gamma}_n^{(g)} = \overline{\Gamma}_n$ be the subset of Γ_n consisting of those cones σ such that $\sigma \cap \Omega_g \neq \varnothing$. For each $\sigma \in \overline{\Gamma}_n$, let ω_σ be an orientation of σ, and for each $\sigma \in \overline{\Gamma}_{n-1}$ define $\delta(\sigma', \sigma)$ as before in (4). With this notation, we can now define the Voronoi complex.

Definition 4.6 The *Voronoi chain complex* $(V_\bullet^{(g)}, d_\bullet)$ is the complex where $V_n^{(g)}$ is the $\mathbb{Q}$–vector space with basis indexed by $\overline{\Gamma}_n$ and the differential $d_n \colon V_n^{(g)} \to V_{n-1}^{(g)}$ is given by

$$d(e_\sigma) := \sum_{\sigma' \in \overline{\Gamma}_{n-1}} \delta(\sigma', \sigma) e_{\sigma'}.$$

Example 4.7 The $\mathrm{GL}_2(\mathbb{Z})$–orbits of alternating cones in Σ_2^{P} are all contained in $\Omega_2^{\mathrm{rt}} \setminus \Omega_2$. Hence the Voronoi complex $V_\bullet^{(2)}$ is zero in all degrees.

There is a natural surjection of chain complexes $P_\bullet^{(g)} \twoheadrightarrow V_\bullet^{(g)}$ given by quotienting $P_n^{(g)}$ by the subcomplex spanned by those cones contained in $\Omega_g^{\mathrm{rt}} \setminus \Omega_g$. Our next goal, achieved in Theorem 4.13, is to show that the kernel of the map above can naturally be identified with $P_\bullet^{(g-1)}$. We begin by noting the following two lemmas studying those cones lying in $\Omega_g^{\mathrm{rt}} \setminus \Omega_g$.

Lemma 4.8 *Let* $\sigma = \mathbb{R}_{\geq 0} \langle v_1 v_1^t, \ldots, v_n v_n^t \rangle$ *with* $v_1, \ldots, v_n \in \mathbb{Z}^g$ *be a perfect cone. Then* σ *is contained in* $\Omega_g^{\mathrm{rt}} \setminus \Omega_g$ *if and only if* $\dim \mathrm{span}_{\mathbb{R}} \langle v_1, v_2, \ldots, v_n \rangle < g$.

Lemma 4.9 *If* $\sigma \in \Sigma_g^{\mathrm{P}}$ *is a perfect cone and* $\sigma \subset \Omega_g^{\mathrm{rt}} \setminus \Omega_g$ *then there is a matrix* $A \in \mathrm{GL}_g(\mathbb{Z})$ *and a cone* $\sigma' \in \Sigma_{g'}^{\mathrm{P}}$, *where* $g' < g$ *and* $\sigma' \cap \Omega_{g'} \neq \varnothing$, *with*

$$(5) \qquad A\sigma A^t = \left\{ \left(\begin{array}{c|c} Q' & 0 \\ \hline 0 & 0 \end{array} \right) \;\middle|\; Q' \in \sigma' \right\}.$$

In this situation, say σ' *is a **reduction** of* σ.

We now show that, in a sense that we shall make precise, the action of $\mathrm{GL}_g(\mathbb{Z})$ on σ does not depend on the ambient matrix size g. For example, given a cone $\sigma \in \Sigma_g^{\mathrm{P}}$ and a reduction $\sigma' \in \Sigma_{g'}^{\mathrm{P}}$ of σ, we will see that σ is alternating if and only if σ' is alternating. We begin with the following definition.

Definition 4.10 Given perfect cones $\sigma_1, \sigma_2 \in \Sigma_g^{\mathrm{P}}$, let $\mathrm{Hom}_{\Omega_g^{\mathrm{rt}}}(\sigma_1, \sigma_2)$ denote the set of morphisms $\rho \colon \sigma_1 \to \sigma_2$ which are restrictions from the action of $\mathrm{GL}_g(\mathbb{Z})$ on Ω_g^{rt}:

$$
\begin{array}{ccc}
\Omega_g^{\mathrm{rt}} & \xrightarrow{\;X \mapsto AXA^t\;} & \Omega_g^{\mathrm{rt}} \\
\uparrow & & \uparrow \\
\sigma_1 & \xrightarrow{\quad \rho \quad} & \sigma_2
\end{array}
$$

The following two results concerning homomorphisms of cones contained in the boundary of Ω_g^{rt} are standard and possibly well known to experts. We include proofs here, however, as we are unaware of suitable references.

Proposition 4.11 *If $\sigma_1, \sigma_2 \in \Sigma_g^P$ are perfect cones contained in $\Omega_g^{\mathrm{rt}} \setminus \Omega_g$ and $\sigma_1', \sigma_2' \in \Sigma_{g'}^P$ are reductions of σ_1 and σ_2 respectively, then there exists a bijection*

$$\mathrm{Hom}_{\Omega_g^{\mathrm{rt}}}(\sigma_1, \sigma_2) \overset{\sim}{\longleftrightarrow} \mathrm{Hom}_{\Omega_{g'}^{\mathrm{rt}}}(\sigma_1', \sigma_2').$$

Proof By Lemma 4.9, we may assume σ_i are in the form of (5). Then if $\rho' \in \mathrm{Hom}_{\Omega_{g'}^{\mathrm{rt}}}(\sigma_1', \sigma_2')$ arises from the action of a matrix $A' \in \mathrm{GL}_{g'}(\mathbb{Z})$ on $\Omega_{g'}^{\mathrm{rt}}$, then extending it by a $(g-g') \times (g-g')$ identity matrix gives a matrix $A \in \mathrm{GL}_g(\mathbb{Z})$ that induces a cone morphism $\rho \colon \sigma_1 \to \sigma_2$.

In the other direction, suppose that $\rho \in \mathrm{Hom}_{\Omega_g^{\mathrm{rt}}}(\sigma_1, \sigma_2)$ comes from the action of a matrix $A \in \mathrm{GL}_g(\mathbb{Z})$ on Ω_g^{rt}. Write $\mathbb{R}^{g'}$ for the coordinate subspace of $\mathbb{R}^g$ of vectors in which the last $g-g'$ coordinates are zero. Let $\sigma_1 = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_n v_n^t \rangle$. By Lemma 4.8, the vectors $v_1, \ldots, v_n$ span $\mathbb{R}^{g'}$. Since $A\sigma_1 A^t = \sigma_2$, it follows again from Lemma 4.8 that $Av_1, \ldots, Av_n$ also span $\mathbb{R}^{g'}$; thus A restricts to a map $A' \colon \mathbb{R}^{g'} \to \mathbb{R}^{g'}$, with $A(\mathbb{Z}^{g'}) \subseteq \mathbb{Z}^{g'}$. Similarly, A^{-1} restricts to $(A')^{-1} \colon \mathbb{R}^{g'} \to \mathbb{R}^{g'}$, and $(A')^{-1}(\mathbb{Z}^{g'}) \subseteq \mathbb{Z}^{g'}$. Therefore $A' \in \mathrm{GL}_{g'}(\mathbb{Z})$ is an invertible integer matrix, with $A'\sigma_1'(A')^t = \sigma_2'$.

Finally, a direct computation shows that these constructions are mutual inverses. $\square$

As a corollary of Proposition 4.11, the properties of being in the same $\mathrm{GL}_g(\mathbb{Z})$–orbit and being alternating do not depend on g — that is, they are preserved by taking reductions.

Corollary 4.12 *Two perfect cones $\sigma_1, \sigma_2 \subset \Omega_g^{\mathrm{rt}} \setminus \Omega_g$ are in the same $\mathrm{GL}_g(\mathbb{Z})$–orbit if and only if there exists a $g' < g$ and reductions $\sigma_1', \sigma_2' \in \Sigma_{g'}^P$ that are in the same $\mathrm{GL}_{g'}(\mathbb{Z})$–orbit. A perfect cone $\sigma \subset \Omega_g^{\mathrm{rt}} \setminus \Omega_g$ is alternating if and only if there exists a reduction $\sigma' \in \Sigma_{g'}^P$ which is alternating.*

Proof Two perfect cones σ_1 and σ_2 are in the same orbit if and only if $\mathrm{Hom}_{\Omega_g^{\mathrm{rt}}}(\sigma_1, \sigma_2)$ is nonempty. Then the claim follows from Proposition 4.11, since $\mathrm{Hom}_{\Omega_g^{\mathrm{rt}}}(\sigma_1, \sigma_2)$ is nonempty if and only if $\mathrm{Hom}_{\Omega_{g'}^{\mathrm{rt}}}(\sigma_1', \sigma_2')$ is nonempty. Similarly, the proof of Proposition 4.11, applied to $\sigma = \sigma_1 = \sigma_2$, shows that σ has an orientation-reversing automorphism if and only if its reduction σ' does. $\square$

Corollary 4.12 allows us to naturally identify the set of $\mathrm{GL}_g(\mathbb{Z})$–orbits of alternating perfect cones in $\Omega_g^{\mathrm{rt}} \setminus \Omega_g$ with the set of $\mathrm{GL}_{g-1}(\mathbb{Z})$–orbits of alternating perfect cones in $\Omega_{g-1}^{\mathrm{rt}}$. Thus, we have the following theorem.

Theorem 4.13 *We have a short exact sequence of chain complexes*

$$0 \to P_\bullet^{(g-1)} \to P_\bullet^{(g)} \xrightarrow{\pi} V_\bullet^{(g)} \to 0.$$

Proof By construction, the kernel of $\pi \colon P_n^{(g)} \to V_n^{(g)}$ is generated by those basis vectors e_σ where $\sigma \in \Gamma_n^{(g)} \setminus \overline{\Gamma}_n^{(g)}$. (Recall that $\Gamma_n^{(g)}$ denotes a set of representatives of alternating $\mathrm{GL}_g(\mathbb{Z})$–orbits of cones

in Σ_g^P, and $\overline{\Gamma}_n^{(g)}$ denotes the subset of those that meet Ω_g.) By Corollary 4.12, such cones are in bijection with elements of $\Gamma_n^{(g-1)}$. The differentials on $P_\bullet^{(g)}$ and $P_\bullet^{(g-1)}$ are defined in the same fashion, so the result follows. □

Theorem 4.13 reflects the stratification of LA_g^{trop} by the spaces $L\Omega_{g'}/\mathrm{GL}_{g'}(\mathbb{Z})$ for $g' = 1, \ldots, g$, which are rational classifying spaces for $\mathrm{GL}_{g'}(\mathbb{Z})$; this is the underlying geometric reason that it is possible to relate the cohomology of $\mathcal{A}_g$ to that of $\mathrm{GL}_{g'}(\mathbb{Z})$, as we do here. This possible relationship was suggested in the more general setting of arithmetic quotients of Hermitian symmetric domains in [40, Section 2.4].

5 The inflation complex and the coloop complex

In this section, we define a subcomplex of $P_\bullet^{(g)}$, called the inflation complex $I_\bullet^{(g)}$. We shall show in Theorem 5.15 that $I_\bullet^{(g)}$ is acyclic. This acyclicity result implies a vanishing result for $H_k(P_\bullet^{(g)})$ in low degrees, obtained in Corollary 5.16, and it is invoked in the computations in the next section for $g = 6$ and $g = 7$. In Section 5.2, we define an analogous subcomplex, the *coloop* complex, of the regular matroid complex, and prove an analogous acyclicity result. The acyclicity of the coloop complex will not be used in this paper, but should likely be useful for future study of the regular matroid complex.

5.1 The inflation complex

Definition 5.1   (i) Let $S \subset \mathbb{Z}^g$ be a finite set. Say $v \in S$ is a $\mathbb{Z}^g$–*coloop* of S if v is part of a $\mathbb{Z}$–basis $v, w_2, \ldots, w_g$ for $\mathbb{Z}^g$ such that any $w \in S \setminus \{v\}$ is in the $\mathbb{Z}$–linear span of $w_2, \ldots, w_g$. Equivalently, v is a $\mathbb{Z}^g$–coloop if, up to the action of $\mathrm{GL}_g(\mathbb{Z})$, we may write $v = (0, \ldots, 0, 1)$ and $w = (*, \ldots, *, 0)$ for all $w \in S \setminus \{v\}$.

   (ii) Now let $\sigma = \sigma[Q]$ be a perfect cone in Σ_g^P. Recall that the set $M(Q)$ of minimal vectors has the property that $v \in M(Q)$ if and only if $-v \in M(Q)$; let $M'(Q) = \{v_1, \ldots, v_n\}$ be a choice of one of $\{v, -v\}$ for each $v \in M(Q)$. So

$$\sigma = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_n v_n^t \rangle.$$

We say $v \in M'(Q)$ is a *coloop* of σ if v is a $\mathbb{Z}^g$–coloop in $M'(Q)$.

Remark 5.2 The definition of a $\mathbb{Z}^g$–coloop is inspired by the notion of a coloop of a matroid (ie an element not belonging to any circuit). Indeed, if $S \subset \mathbb{Z}^g$ is any finite set and $v \in S$, then v being a $\mathbb{Z}^g$–coloop of S implies that v is a coloop of S, considered as vectors in $\mathbb{R}^g$. The converse does not hold: for example, let $(v_1, v_2) = ((0, 1), (3, 2))$. Then v_1 and v_2 are coloops of the matroid $M(v_1, v_2)$ over $\mathbb{R}$, but neither is a $\mathbb{Z}^2$–coloop.

On the other hand, we prove in Lemma 5.22 that if M is a regular matroid, then M has a coloop if and only if a totally unimodular matrix A representing M has column vectors with a $\mathbb{Z}^g$–coloop, if and only if $\sigma(M)$ has a coloop in the sense of Definition 5.1(ii).

Example 5.3   Consider the quadratic form defined by the positive definite matrix

$$Q = \begin{pmatrix} 2 & \frac{1}{2} & 1 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ 1 & \frac{1}{2} & 1 \end{pmatrix}.$$

The minimum of Q on $\mathbb{Z}^3 - \{0\}$ is 1 and

$$M(Q) = \{(0, \pm 1, 0), (0, 0, \pm 1), \pm(1, 0, -1), \pm(0, 1, -1)\} \subset \mathbb{R}^3.$$

The corresponding perfect cone $\sigma[Q]$ has a coloop, in particular, letting

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{pmatrix}$$

we see that

$$A \begin{pmatrix} 2 & \frac{1}{2} & 1 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ 1 & \frac{1}{2} & 1 \end{pmatrix} A^t = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and so $M(AQA^t)$ is $\{(\pm 1, 0, 0), (0, \pm 1, 0), \pm(1, 1, 0), (0, 0, \pm 1)\}$.

The cone $\sigma[Q]$ can also be realized as the matroidal cone $\sigma[M(G)]$ where G is the following graph:
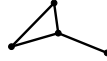
The coloop corresponds to the bridge edge of G.

Lemma 5.4   *Let $S = \{v_1, \ldots, v_n\} \subset \mathbb{Z}^g$ and suppose that $v_1 \neq v_2$ are both $\mathbb{Z}^g$–coloops for S. Then there is a $\mathbb{Z}$–basis $v_1, v_2, w_3, \ldots, w_g$ for $\mathbb{Z}^g$ such that*

$$v_3, \ldots, v_n \in \mathbb{Z}\langle w_3, \ldots, w_g \rangle.$$

Proof   By restricting to $\mathbb{R}\langle v_1, \ldots, v_n \rangle$, we may assume that $\mathbb{R}\langle v_1, \ldots, v_n \rangle = \mathbb{R}^g$. Now the fact that both v_1 and v_2, being coloops, are in every basis for $\mathbb{R}^g$ chosen from $\{v_1, \ldots, v_n\}$, implies that $v_3, \ldots, v_n$ span a $(g-2)$–dimensional subspace V of $\mathbb{R}^g$. Let $w_3, \ldots, w_g$ be a $\mathbb{Z}$–basis for $V \cap \mathbb{Z}^g$. We need only verify that $v_1, v_2, w_3, \ldots, w_g$ form a $\mathbb{Z}$–basis for $\mathbb{Z}^g$. Let $x \in \mathbb{Z}^g$. Since $v_1, v_2, w_3, \ldots, w_g$ form a $\mathbb{Q}$–basis, we have

$$x = a_1 v_1 + a_2 v_2 + a_3 w_3 + \cdots + a_g w_g \quad \text{for some } a_i \in \mathbb{Q},$$

and it suffices to show that $a_i \in \mathbb{Z}$ for all $i = 1, 2, 3, \ldots, g$. First, we show $a_1 \in \mathbb{Z}$. Since v_1 is a $\mathbb{Z}^g$–coloop, after a change of $\mathbb{Z}$–basis, we may assume that $v_1 = (0, \ldots, 0, 1)$ and that $v_2, \ldots, v_n$ have last coordinate zero. Since $w_3, \ldots, w_g \in \mathrm{span}_{\mathbb{R}}\langle v_3, \ldots, v_n \rangle$, each w_i also has last coordinate zero. Therefore $a_1 \in \mathbb{Z}$. By a similar argument, $a_2 \in \mathbb{Z}$. Then $a_3 w_3 + \cdots + a_g w_g \in V \cap \mathbb{Z}^g$ for $a_i \in \mathbb{Q}$. But $w_3, \ldots, w_g$ is a $\mathbb{Z}$–basis for $V \cap \mathbb{Z}^g$. Therefore it must be that $a_3, \ldots, a_g \in \mathbb{Z}$, as desired. $\square$

Corollary 5.5 *Let $S \subset \mathbb{Z}^{g-1}$ be a finite set, and identify $\mathbb{Z}^{g-1}$ with its image in $\mathbb{Z}^g$ under $w \mapsto (w, 0)$. Then v is a coloop of S if and only if v is a coloop of $S \cup \{e_g\} \subset \mathbb{Z}^g$.*

Proof The forward direction is direct from the definitions. For the backward direction, suppose v is a coloop of $S \cup \{e_g\}$. Then *both* v and e_g are coloops, so by Lemma 5.4, up to the action of $\mathrm{GL}_g(\mathbb{Z})$, we may assume that $v = e_{g-1}$ and $w = (*, \ldots, *, 0, 0)$ for all $w \in S \setminus \{v\}$. Therefore v was a coloop of $S \subset \mathbb{Z}^{g-1}$. □

Corollary 5.6 *A cone with two or more coloops is not alternating. That is, if $\sigma = \sigma[Q]$, where $v \neq v' \in M'(Q)$ are two distinct coloops, then σ is not alternating.*

Proof The cone σ has an orientation-reversing automorphism induced by an element of $\mathrm{GL}_g(\mathbb{Z})$ swapping the two coloops. □

We now describe two operations on cones, inflation and deflation, which add or remove coloops, respectively. Inflation is described in [22, Section 6.1], and can be performed for any cone, but we shall consider it for the cones in the set $\Sigma^P_{g,\mathrm{nco}}[n]$ defined below.

Recall that $\Sigma^P_g[n]$ denotes the set of $(n+1)$–dimensional perfect cones, and $\Sigma^P_g[n]/\mathrm{GL}_g(\mathbb{Z})$ denotes the collection of $\mathrm{GL}_g(\mathbb{Z})$–orbits of $(n+1)$–dimensional perfect cones.

Definition 5.7 We define two subsets of $\Sigma^P_g[n]$ as follows:

$$\Sigma^P_{g,\mathrm{nco}}[n] := \{\sigma \in \Sigma^P_g[n] : \mathrm{rank}(\sigma) \leq g - 1 \text{ and } \sigma \text{ has no coloop}\},$$
$$\Sigma^P_{g,\mathrm{co}}[n] := \{\sigma \in \Sigma^P_g[n] : \sigma \text{ has exactly one coloop}\}.$$

We then define $\Sigma^P_{g,\mathrm{nco}}[n]/\mathrm{GL}_g(\mathbb{Z})$ and $\Sigma^P_{g,\mathrm{co}}[n]/\mathrm{GL}_g(\mathbb{Z})$ to be the collection of $\mathrm{GL}_g(\mathbb{Z})$–orbits of the respective sets.

We now define inflation and deflation as operations on $\Sigma^P_{g,\mathrm{nco}}[n]/\mathrm{GL}_g(\mathbb{Z})$ and $\Sigma^P_{g,\mathrm{co}}[n]/\mathrm{GL}_g(\mathbb{Z})$, and we show these operations are well defined in Lemma 5.12.

Definition 5.8 *Inflation* is the map

$$\mathrm{ifl} \colon \Sigma^P_{g,\mathrm{nco}}[n]/\mathrm{GL}_g(\mathbb{Z}) \to \Sigma^P_{g,\mathrm{co}}[n+1]/\mathrm{GL}_g(\mathbb{Z})$$

defined as follows. Given an element of $\Sigma^P_{g,\mathrm{nco}}[n]/\mathrm{GL}_g(\mathbb{Z})$, choose a representative

$$\sigma = \mathbb{R}_{\geq 0}\langle w_1 w_1^t, \ldots, w_k w_k^t \rangle$$

so that the gth entry of each w_i is zero; see Lemma 4.9. Let

$$\widetilde{\sigma} = \mathbb{R}_{\geq 0}\langle w_1 w_1^t, \ldots, w_k w_k^t, e_g e_g^t \rangle.$$

Then set $\mathrm{ifl}([\sigma]) = [\widetilde{\sigma}]$.

Remark 5.9 We check that inflation is well defined in Lemma 5.12. However, we pause to point out that $\widetilde{\sigma}$ is indeed a perfect cone, as noted in [22, Section 6.1]: if $Q \in \Omega_{g-1}$ is a positive definite quadratic form such that $\sigma[Q] = \mathbb{R}_{\geq 0}\langle \widetilde{w}_1 \widetilde{w}_1^t, \ldots, \widetilde{w}_k \widetilde{w}_k^t \rangle$, where $\widetilde{w}_i$ denotes the truncation of w_i by the last entry, then the inflation of σ is the cone associated to the quadratic form

$$\widetilde{Q} = \left(\begin{array}{c|c} Q & 0 \\ \hline 0 & m(Q) \end{array} \right),$$

where $m(Q)$ is the minimum value of Q on $\mathbb{Z}^{g-1} \setminus \{0\}$. Moreover, $\widetilde{\sigma}$ has exactly one coloop by Corollary 5.5 and the fact that σ had no coloops.

Example 5.10 Continuing Example 5.3, we see that if Q' is the positive definite quadratic form

$$Q' = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 2 \end{pmatrix},$$

then $M(Q') = \{\pm(1,0,0), \pm(0,1,0), \pm(1,-1,0)\}$. Thus, the cone $\sigma[Q]$ is the inflation of $\sigma[Q']$. We may describe the cone $\sigma[Q']$ as

$$\sigma[Q'] = \left\{ \left(\begin{array}{c|c} Q'' & 0 \\ \hline 0 & 0 \end{array} \right) \ \Big| \ Q'' \in \sigma\left[\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \right] \right\},$$

from which we see that $\sigma[Q']$ does not meet Ω_3, but its inflation $\sigma[Q]$ does. In general, the inflation of a perfect cone corresponding to a quadratic form of rank r will itself be a perfect cone corresponding to a quadratic form of rank $r+1$.

Definition 5.11 We define the *deflation* operation as a map

$$\mathrm{dfl}\colon \Sigma^P_{g,\mathrm{co}}[n+1]/\mathrm{GL}_g(\mathbb{Z}) \to \Sigma^P_{g,\mathrm{nco}}[n]/\mathrm{GL}_g(\mathbb{Z})$$

given as follows. Given an element of $\Sigma^P_{g,\mathrm{co}}[n+1]/\mathrm{GL}_g(\mathbb{Z})$, pick a $\mathrm{GL}_g(\mathbb{Z})$–representative

$$\widetilde{\sigma} = \mathbb{R}_{\geq 0}\langle w_1 w_1^t, \ldots, w_k w_k^t, e_g e_g^t \rangle,$$

where each w_i is zero in the last coordinate. Let $\sigma = \mathbb{R}_{\geq 0}\langle w_1 w_1^t, \ldots, w_k w_k^t \rangle$. It is routine to check that σ really is a perfect cone, and moreover, it has no coloops by Corollary 5.5. Then we set $\mathrm{dfl}([\widetilde{\sigma}]) = [\sigma]$. We now show that inflation and deflation are well defined.

Lemma 5.12 *For each $n \in \mathbb{N}$, inflation is a well-defined operation on $\Sigma^P_{g,\mathrm{nco}}[n]/\mathrm{GL}_g(\mathbb{Z})$, and deflation is a well-defined operation on $\Sigma^P_{g,\mathrm{co}}[n+1]/\mathrm{GL}_g(\mathbb{Z})$. Furthermore, these operations are inverses of each other.*

Proof We start with inflation. Given $[\sigma] \in \Sigma^P_{g,\mathrm{nco}}[n]/\mathrm{GL}_g(\mathbb{Z})$, let $\sigma_1 = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_k v_k^t \rangle$ and $\sigma_2 = \mathbb{R}_{\geq 0}\langle w_1 w_1^t, \ldots, w_k w_k^t \rangle$ be two $\mathrm{GL}_g(\mathbb{Z})$–representatives of $[\sigma]$ such that the gth entry of each

of the v_i, w_j is zero. By Proposition 4.11, there exist reductions σ_1' of σ_1 and σ_2' of σ_2 as well as an $A \in \mathrm{GL}_{g-1}(\mathbb{Z})$ sending σ_1' to σ_2' in $\Omega_{g-1}^{\mathrm{rt}}$. Then

$$A' = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & 1 \end{array}\right)$$

yields an equivalence between the two inflations.

Now let $[\sigma] \in \Sigma_{g,\mathrm{co}}^{\mathrm{P}}[n+1]/\mathrm{GL}_g(\mathbb{Z})$. Let

$$\sigma_1 = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_k v_k^t, e_g e_g^t\rangle, \ \sigma_2 = \mathbb{R}_{\geq 0}\langle w_1 w_1^t, \ldots, w_k w_k^t, e_g e_g^t\rangle$$

be two $\mathrm{GL}_g(\mathbb{Z})$ representatives of $[\sigma]$ such that the v_i, w_j have g^{th} coordinate zero. Then by the proof of Proposition 4.11, there exists an $A \in \mathrm{GL}_g(\mathbb{Z})$ such that

$$Av_i = \pm w_i \quad \text{for } i = 1, \ldots, k,$$

$$Ae_g = \pm e_g,$$

possibly after reordering the v_i. Indeed, A must take the coloop $\pm e_g$ to $\pm e_g$. Then A gives an equivalence between the deflations $\mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_k v_k^t\rangle \sim \mathbb{R}_{\geq 0}\langle w_1 w_1^t, \ldots, w_k w_k^t\rangle$.

We now have that inflation and deflation are well defined, and it is clear from the definitions that these two operations are inverses. $\qquad\square$

Lemma 5.13 *Let $\sigma = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_k v_k^t\rangle$ be a perfect cone in Σ_g^{P} of rank $< g$ with no coloop. Then $[\sigma]$ is alternating if and only if* $\mathrm{ifl}([\sigma])$ *is alternating.*

Proof We may assume that $v_1, \ldots, v_n$ have last coordinate 0, so that, letting

$$\tilde{\sigma} = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \ldots, v_k v_k^t, e_g e_g^t\rangle,$$

we have $\mathrm{ifl}(\sigma) = \tilde{\sigma}$. We claim there is a natural bijection

$$(6) \qquad\qquad\qquad \mathrm{Aut}(\sigma) \longleftrightarrow \mathrm{Aut}(\tilde{\sigma}),$$

where $\mathrm{Aut}(\sigma) = \mathrm{Hom}_{\Omega_g^{\mathrm{rt}}}(\sigma, \sigma)$ (see Definition 4.10) and similarly for $\mathrm{Aut}(\tilde{\sigma})$. Moreover, we claim that (6) takes orientation-preserving/reversing automorphisms of σ to orientation-preserving/reversing automorphisms of $\tilde{\sigma}$, respectively.

Given $\rho \in \mathrm{Aut}(\sigma)$ arising from a matrix $A \in \mathrm{GL}_{g-1}(\mathbb{Z})$, the matrix

$$\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix}$$

yields an automorphism $\tilde{\rho}$ of $\tilde{\sigma}$. The linear span of $\tilde{\sigma}$ is the sum of the linear span of σ and that of $e_g e_g^t$. Moreover, $\tilde{\rho}$ fixes the ray $e_g e_g^t$ of $\tilde{\sigma}$ and acts on the linear span of σ according to A; in particular, $\tilde{\rho}$ is orientation-preserving if and only if ρ was.

Next, suppose $\widetilde{A} \in \mathrm{GL}_g(\mathbb{Z})$ induces $\widetilde{\rho} \in \mathrm{Aut}(\widetilde{\sigma})$. Recall that e_g is the only coloop of $\widetilde{\sigma}$, by Corollary 5.5. Therefore $Ae_g = \pm e_g$, and hence A induces an automorphism of σ. Finally, it is routine to check that the maps constructed between $\mathrm{Aut}(\sigma)$ and $\mathrm{Aut}(\widetilde{\sigma})$ are two-sided inverses.   □

Definition 5.14 Let $I_\bullet^{(g)}$ be the subcomplex of $P_\bullet^{(g)}$ which is generated in degree n by cones $\sigma \in \Gamma_n$ of rank $\leq g - 1$ and cones of rank g with a coloop.

Theorem 5.15 *The chain complex $I_\bullet^{(g)}$ is acyclic.*

Proof By Lemmas 5.12 and 5.13 there is a matching of cones generating $I_\bullet^{(g)}$, given by

$$\sigma \to \begin{cases} \mathrm{ifl}(\sigma) & \text{if } \sigma \text{ has no coloop,} \\ \mathrm{dfl}(\sigma) & \text{if } \sigma \text{ has a coloop.} \end{cases}$$

Here, we have abused notation slightly, since ifl is an operation on orbits rather than orbit representatives. Thus, when we write $\mathrm{ifl}(\sigma) = \sigma'$ for $\sigma \in \Gamma_n$, we mean that σ' is the unique orbit representative in Γ_{n+1} such that $\mathrm{ifl}([\sigma]) = [\sigma']$. Similarly for deflation.

Now let σ be a generator in $I_\bullet^{(g)}$ of maximal degree; then σ must have a coloop v. We claim that $\sigma' = \mathrm{dfl}(\sigma)$ is not a facet of any other generator $\tau \neq \sigma$ of $I_\bullet^{(g)}$. Indeed, suppose that

$$\tau = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, \dots, v_k v_k^t \rangle$$

is a generator of $I_\bullet^{(g)}$ containing σ' as a facet. If τ had a coloop, say v_n, then since σ' has no coloop, σ' must not contain the ray $v_n v_n^t$. But

$$\mathbb{R}_{\geq 0}\langle v_1 v_1^t, \dots, v_{k-1} v_{k-1}^t \rangle$$

is already a facet of τ, so it must be $\sigma' = \mathrm{dfl}(\sigma) = \mathrm{dfl}(\tau)$, implying $\tau = \sigma$. So τ has no coloop. But then $\mathrm{ifl}(\tau)$ is a generator of $I_\bullet^{(g)}$ and it would have higher rank than σ.

Thus, the complex $I_\bullet^{(g)'}$ spanned by all cones except σ and $\mathrm{dfl}(\sigma)$ is a subcomplex. Then we have a short exact sequence

$$0 \to I_\bullet^{(g)'} \to I_\bullet^{(g)} \to I_\bullet^{(g)}/I_\bullet^{(g)'} \to 0,$$

where $I_\bullet^{(g)}/I_\bullet^{(g)'}$ is isomorphic to $0 \to \sigma \to \mathrm{dfl}(\text{œ}) \to 0$. Hence $I_\bullet^{(g)'} \to I_\bullet^{(g)}$ is a quasi-isomorphism. Repeating this, we deduce inductively that I is quasi-isomorphic to 0.   □

As a corollary of Theorem 5.15, we are able to prove the following vanishing result for the cohomology of $P_\bullet^{(g)}$ in low degrees.

Corollary 5.16 *If $k \leq g - 2$, then $H_k(P_\bullet^{(g)}) = 0$.*

Proof Since the inflation complex $I_\bullet^{(g)}$ is acyclic by Theorem 5.15, it is enough to show that $I_k^{(g)} = P_k^{(g)}$ for all $k \leq g - 2$. For this, we simply need the well-known fact that the rank of a perfect cone is at most

its dimension. Indeed, let $\sigma = \mathbb{R}_{\geq 0}\langle v_1 v_1^t, v_2 v_2^t, \ldots, v_n v_n^t \rangle \in \Sigma_g^{\mathrm{P}}$ be an alternating cone of dimension $k+1$. If $Q \in \sigma$ then

$$Q = \lambda_1 v_{i_1} v_{i_1}^t + \lambda_2 v_{i_2} v_{i_2}^t + \cdots + \lambda_{k+1} v_{i_{k+1}} v_{i_{k+1}}^t$$

for some $\{i_1, \ldots, i_{k+1}\} \subset \{1, \ldots, n\}$ and some $\lambda_i \in \mathbb{R}_{\geq 0}$. In particular, since $v_j v_j^t$ is a rank-one quadratic form, this implies that the rank of Q is at most $k+1$. Thus, if $k+1 \leq g-1$, then $\mathrm{rank}(\sigma) < g$, implying that the orbit of σ represents an element of $I_k^{(g)}$. □

Remark 5.17 The inflation operation is, of course, a special case of taking the block sum of two perfect cones. In this way one obtains a product map on chain complexes $P^{(g_1)} \otimes P^{(g_2)} \to P^{(g_1+g_2)}$, and a corresponding product on homology. This is reminiscent of the result of [25] describing the stable cohomology of the matroidal partial compactifications $\overline{\mathcal{A}}_g^{\mathrm{matr}}$ via 1–sums of irreducible regular matroids. Perhaps if one had nonvanishing statements for the latter product, then the cohomology classes detected in this paper could be used to construct infinite families of top-weight classes in $\mathcal{A}_g$.

Remark 5.18 The proof of Corollary 5.16 shows that any cone of dimension less than or equal to $g-1$ does not intersect Ω_g. This implies that $H_k(V_\bullet^{(g)}) = 0$ for all $k \leq g-2$.

Remark 5.19 The virtual cohomological dimension of $\mathcal{A}_g$ is

$$\mathrm{vcd}(\mathcal{A}_g) = \mathrm{vcd}(\mathrm{Sp}(2g, \mathbb{Z})) = g^2$$

by [7]; see [15]. In particular,

$$\mathrm{Gr}_{g^2+g}^W H^i(\mathcal{A}_g; \mathbb{Q}) = 0 \quad \text{for all } i > g^2,$$

which is equivalent, setting $i = 2\dim(\mathcal{A}_g) - j - 1 = g^2 + g - j - 1$, to

$$H_j(P^{(g)}) = 0 \quad \text{for all } j < g-1.$$

Corollary 5.16 thus reproves, in a completely different way, the vanishing in top weight of rational cohomology of $\mathcal{A}_g$ in degree above the virtual cohomological dimension.

5.2 The regular matroid complex and inflation

In this section, we introduce two combinatorially defined subcomplexes $R_\bullet^{(g)}$ and $C_\bullet^{(g)}$ of $P_\bullet^{(g)}$, coming from regular matroids and regular matroids with coloops, respectively. These are not used further in this paper. Nevertheless, the matroidal cones in Σ_g^P have geometric significance: Alexeev and Brunyate, in proving the existence of a compactified Torelli map $\overline{\mathcal{M}}_g \to \overline{\mathcal{A}}_g^{\mathrm{perf}}$, conjectured an open locus on which $\overline{\mathcal{A}}_g^{\mathrm{perf}}$ and $\overline{\mathcal{A}}_g^{\mathrm{Vor}}$ are isomorphic and on which the two Torelli maps $\overline{\mathcal{M}}_g \to \overline{\mathcal{A}}_g^{\mathrm{perf}}$ and $\overline{\mathcal{M}}_g \to \overline{\mathcal{A}}_g^{\mathrm{Vor}}$ agree [2]. The fourth author and Viviani [37] verified their conjecture, showing that the *matroidal partial compactification* $\mathcal{A}_g^{\mathrm{matr}}$, whose strata correspond to cones arising from regular matroids, is the largest such open subset. For possible future use in studying $R^{(g)}$, we establish in this section that the complex $C_\bullet^{(g)}$, which is a matroid analogue of $I_\bullet^{(g)}$, is acyclic.

Given a cone $\sigma \in \Sigma_g^P[n]$, we say that σ is a matroidal cone if and only if there exists a simple, regular matroid M of rank at most g such that $[\sigma] = \sigma(M)$, where $\sigma(M)$ is as described in Construction 2.12. Matroidal cones are simplicial. Since the faces of a matroidal cone are themselves matroidal cones, the set of representatives of alternating cones arising from simple, regular matroids forms a subcomplex of $P^{(g)}$.

Definition 5.20 The *regular matroid complex* $R_\bullet^{(g)}$ is the subcomplex of $P_\bullet^{(g)}$ generated in degree n by cones $\sigma \in \Gamma_n$ such that σ is a matroidal cone.

Remark 5.21 When $g = 2$ and $g = 3$, the complexes $R_\bullet^{(g)}$ and $P_\bullet^{(g)}$ are in fact equal. It would be interesting to understand in general how much larger $P_\bullet^{(g)}$ is compared to $R_\bullet^{(g)}$.

Recall that an element e of a matroid M is a *coloop* if it does not belong to any of the circuits of M; equivalently, e is a coloop if it belongs to every base of M. When M is a regular matroid, this is equivalent to the existence of a totally unimodular matrix $A = [v_1, v_2, \ldots, v_n]$ representing M such that $v_i = (*, *, \ldots, *, 0)$ for $i = 1, 2, \ldots, n-1$ and $e = v_n = (0, 0, \ldots, 0, 1)$. It is worth establishing that the notions of a matroid coloop and a $\mathbb{Z}^g$–coloop agree for matroidal cones, as we show in the next lemma.

Lemma 5.22 *Let M be a simple, regular matroid of rank $\leq g$. The cone $\sigma(M)$ has a $\mathbb{Z}^g$–coloop if and only if the matroid M has a coloop.*

Proof Suppose that $\sigma(M)$ has a $\mathbb{Z}^g$–coloop. By definition, there exists a quadratic form $Q \in \Omega_g$ such that $[\sigma(Q)] = \sigma(M)$ and $M'(Q) = \{v_1, v_2, \ldots, v_n\}$ where $v_i = (*, *, \ldots, *, 0)$ for $i = 1, \ldots, n-1$ and $v_n = (0, 0, \ldots, 0, 1)$. Then by the construction of $\sigma(M)$, the matrix $A = [v_1, v_2, \ldots, v_n]$ is a totally unimodular matrix representing M over $\mathbb{R}$. Therefore v_n is a coloop of the matroid M.

For the other direction, suppose that the regular matroid M on the ground set $\{1, \ldots, n\}$ is represented by a full-rank totally unimodular $g' \times n$ matrix $A = [v_1, v_2, \ldots, v_n]$ for some $g' \leq g$, and that n is a coloop of M. Then n is in every base of M, so v_n is in every full-rank $g' \times g'$ submatrix of A. Reorder so that the rightmost $g' \times g'$ submatrix is full rank; call it B. Then $B \in \mathrm{GL}_{g'}(\mathbb{Z})$ by total unimodularity of A. Consider the matrix $B^{-1}A$, which still represents M. The rightmost $g' \times g'$ submatrix of $B^{-1}A$ is the identity. Moreover, each of the first $n - g'$ columns is of the form $(*, \ldots, *, 0)$, for otherwise it could replace the last column in the rightmost square submatrix to form a full-rank square matrix, contradicting that n was a coloop. This shows that v_n is a $\mathbb{Z}^{g'}$–coloop of $v_1, \ldots, v_n \in \mathbb{Z}^{g'}$, and after padding by zeroes, v_n is a $\mathbb{Z}^g$–coloop of the of $v_1, \ldots, v_n$. $\square$

Definition 5.23 The *coloop complex* $C_\bullet^{(g)}$ is the subcomplex of $P_\bullet^{(g)}$ generated in degree n by cones $\sigma \in \Gamma_n$ such that σ is a matroidal cone and either

  (i)   the rank of σ is $< g$, or

  (ii) the rank of σ is equal to g and σ has one coloop.

By Lemma 5.22, the generators for $C_\bullet^{(g)}$ are the generators of $R_\bullet^{(g)}$ that are also generators of $I_\bullet^{(g)}$; in summary, we have inclusions of complexes

$$
\begin{array}{ccc}
C_\bullet^{(g)} & \hookrightarrow & R_\bullet^{(g)} \\
\Big\uparrow & & \Big\uparrow \\
\Big\downarrow & & \Big\downarrow \\
I_\bullet^{(g)} & \hookrightarrow & P_\bullet^{(g)}
\end{array}
$$

Similar to the inflation complex, the coloop complex is acyclic.

Theorem 5.24 *The chain complex $C_\bullet^{(g)}$ is acyclic.*

We omit the details of the proof of Theorem 5.24. It is closely analogous to the proof of Theorem 5.15, the key step being the following lemma.

Lemma 5.25 *There is a bijection of sets between*

$$
\left\{ \begin{array}{l} \text{alternating, regular matroids} \\ \text{of rank} < g \text{ with } 0 \text{ coloops} \end{array} \right\} \xleftrightarrow{\ \sim\ } \left\{ \begin{array}{l} \text{alternating, regular matroids} \\ \text{of rank} \leq g \text{ with } 1 \text{ coloop} \end{array} \right\}.
$$

6 Computations on the cohomology of $\mathcal{A}_g$

In this section, we compute the top-weight cohomology of $\mathcal{A}_g$ for $3 \leq g \leq 7$, proving Theorem A. When $g = 3$, 4 and 5, we do this by studying the cones of Σ_g^{P} arising from matroids, from which we explicitly compute the chain complex $P_\bullet^{(g)}$. We handle the cases when $g = 6$ and $g = 7$ by utilizing the long exact sequence in homology arising from Theorem 4.13, as well as the fact that the inflation subcomplex $I_\bullet^{(g)}$ is acyclic; see Theorem 5.15. Additionally, we prove a vanishing result for the top-weight cohomology of $\mathcal{A}_g$ for $g = 8$, 9 and 10 in Theorem 6.16.

6.1 The complex $P_\bullet^{(3)}$

For $g = 3$, the fact that every perfect cone is matroidal allows us to compute the complex $P_\bullet^{(3)}$ directly. Using this description of $P_\bullet^{(3)}$, we then compute the top-weight cohomology of $\mathcal{A}_3$.

Proposition 6.1 *The chain complex $P_\bullet^{(3)}$ is*

$$
\begin{array}{ccccccc}
P_5^{(3)} & P_4^{(3)} & P_3^{(3)} & P_2^{(3)} & P_1^{(3)} & P_0^{(3)} & P_{-1}^{(3)}
\end{array}
$$

$$
0 \longrightarrow \mathbb{Q} \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \mathbb{Q} \xrightarrow{\ \sim\ } \mathbb{Q} \longrightarrow 0.
$$

Proof The only top-dimensional perfect cone of $\Sigma_3^{\mathrm{P}}/\mathrm{GL}_3(\mathbb{Z})$ is the principal cone σ_3^{prin} coming from the complete graph K_4; see [44, page 151]. The principal cone σ_3^{prin} is alternating because the automorphisms of K_4 are all alternating permutations of its edges, and every automorphism of σ_3^{prin} arises from $\mathrm{Aut}(K_4)$ by Remark 2.14. Thus, we have $P_5^{(3)} \cong \mathbb{Q}$.
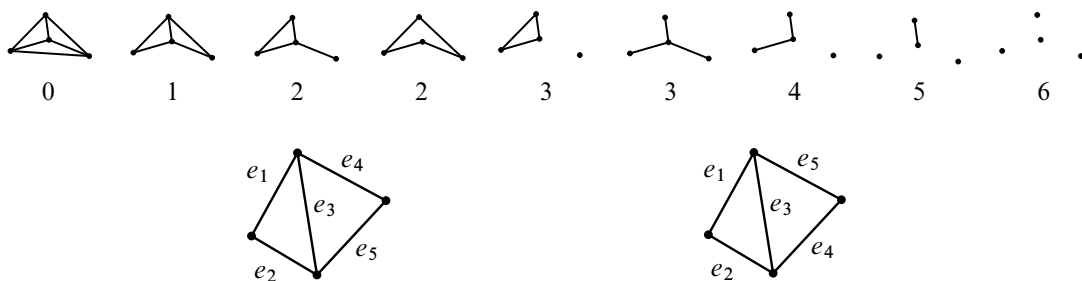
Figure 2: Top: graphs obtained by deleting the indicated number of edges from K_4, giving isomorphism classes of graphic matroids. Bottom: an automorphism of $M[K_4\backslash\{e_6\}]$ interchanging e_4 and e_5.

The automorphisms on the codimension i faces of σ_3^{prin} arise from matroids of graphs obtained from K_4 by deleting i edges; see Figure 2, top. For $i = 1, \ldots, 4$, each of the matroids associated to graphs with i edges removed from K_4 has an automorphism given by an odd permutation of the edges; see Figure 2, bottom, for an example. So we have $P_j^{(3)} = 0$ for $1 \le j \le 4$. The single ray and vertex of $\Sigma_3^{\mathrm{P}}/\mathrm{GL}_3(\mathbb{Z})$ are alternating, so $P_j^{(3)} \cong \mathbb{Q}$ for $j = 0$ and $j = -1$. □

Theorem 6.2 *The top-weight cohomology of $\mathcal{A}_3$ is*

$$\mathrm{Gr}_{12}^W H^i(\mathcal{A}_3; \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } i = 6, \\ 0 & \text{else.} \end{cases}$$

Proof The top-weight cohomology of $\mathcal{A}_3$ is the homology of $P_\bullet^{(3)}$ by Theorem 3.1. □

Remark 6.3 Theorem 6.2 agrees with the work of Hain [27], who computes the full cohomology ring of $\mathcal{A}_3$. Hain deduces in particular $H^6(\mathcal{A}_3; \mathbb{Q}) = E$ where E is a mixed Hodge structure that is an extension $0 \to \mathbb{Q}(-3) \to E \to \mathbb{Q}(-6) \to 0$, where $\mathbb{Q}(n)$ denotes the Tate Hodge structure of dimension one and weight $-2n$.

Example 6.4 While we do not need it here, we note that using the fact that all of the perfect cones in Σ_3^{P} arise from graphic matroids, one can check that the inflation complex $I_\bullet^{(3)}$ is

$$
\begin{array}{ccccccc}
I_5^{(3)} & I_4^{(3)} & I_3^{(3)} & I_2^{(3)} & I_1^{(3)} & I_0^{(3)} & I_{-1}^{(3)}
\end{array}
$$

$$0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \mathbb{Q} \xrightarrow{\ \sim\ } \mathbb{Q} \longrightarrow 0.$$

6.2 The complex $P_\bullet^{(4)}$

In this section, we explicitly compute the complex $P_\bullet^{(4)}$ by using the matroidal description of the principal cone given in Section 2.5 together with the description of a similar complex for $\mathrm{SL}_g(\mathbb{Z})$–alternating cones described in [36]. We then use $P_\bullet^{(4)}$ to compute the top-weight cohomology of $\mathcal{A}_4$.

i	C_i	$\mathrm{SL}_4(\mathbb{Z})$–alternating cones of dim $i+1$	$\mathrm{GL}_4(\mathbb{Z})$–alternating?
4	$\mathbb{Q}$	$\sigma(\text{⬡})$	no
5	$\mathbb{Q}$	$\sigma(\text{△})$	no
6	$\mathbb{Q}$	$\sigma(\text{⬡})$	yes
8	$\mathbb{Q}$	$\sigma(\text{⬡})$	no
9	$\mathbb{Q}^2$	$\sigma_4^{\mathrm{prin}}, \sigma(D_4)$	no

Table 1: $\mathrm{SL}_4(\mathbb{Z})$–alternating cones of $\Sigma_4^P/\mathrm{SL}_4(\mathbb{Z})$.

Proposition 6.5 *The chain complex $P_\bullet^{(4)}$ is*

$$
\begin{array}{ccccccccccc}
P_9^{(4)} & P_8^{(4)} & P_7^{(4)} & P_6^{(4)} & P_5^{(4)} & P_4^{(4)} & P_3^{(4)} & P_2^{(4)} & P_1^{(4)} & P_0^{(4)} & P_{-1}^{(4)}
\end{array}
$$

$$
0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \mathbb{Q} \xrightarrow{\sim} \mathbb{Q} \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \mathbb{Q} \xrightarrow{\sim} \mathbb{Q} \to 0
$$

Proof By Theorem 4.13, we have, in any degree ℓ, that $\dim P_\ell^{(4)} = \dim P_\ell^{(3)} + \dim V_\ell^{(4)}$. We have already computed $P_\bullet^{(3)}$, so we now compute $V_\bullet^{(4)}$. In [36], the authors compute a complex $C_\bullet$ which is generated in degree i by the $(i+1)$–dimensional $\mathrm{SL}_4(\mathbb{Z})$–alternating perfect cones meeting Ω_g up to $\mathrm{SL}_4(\mathbb{Z})$–equivalence. Their results [36, Proposition 3.1] are summarized in the first three columns of Table 1. The cone $\sigma(D_4)$ is the cone corresponding to the quadratic form

$$
D_4 = \begin{bmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}.
$$

For $i \notin \{4,5,6,8,9\}$, they show that $C_i = 0$.

We now compute the Voronoi complex $V_\bullet^{(4)}$. As far as we know, this computation—for $\mathrm{GL}_4(\mathbb{Z})$, as opposed to $\mathrm{SL}_4(\mathbb{Z})$—constitutes a small gap in the literature, which we fill here. To obtain the complex $V_\bullet^{(4)}$, we must pass from $\mathrm{SL}_4(\mathbb{Z})$ to $\mathrm{GL}_4(\mathbb{Z})$. In doing so, two things may happen. First, two $\mathrm{SL}_4(\mathbb{Z})$–inequivalent cones may be $\mathrm{GL}_4(\mathbb{Z})$–equivalent. This does not occur by the corollary following [36, Lemma 4.4]: the $\mathrm{GL}_4(\mathbb{Z})$–orbits of cones in Σ_4^P are equal to the $\mathrm{SL}_4(\mathbb{Z})$–orbits. Second, a cone which is $\mathrm{SL}_4(\mathbb{Z})$–alternating may no longer be $\mathrm{GL}_4(\mathbb{Z})$–alternating. We now check whether this occurs for the cones in Table 1.

In degree 9, neither cone is alternating since transposition matrices stabilize these cones but reverse orientation, as is observed in [36, page 107]. In degrees 4, 5, and 8 the graphic matroids giving rise to each of the cones in Table 1 have an automorphism coming from an odd permutation of the ground set elements, so these cones are not alternating. Therefore $V_9^{(4)} = V_8^{(4)} = V_5^{(4)} = V_4^{(4)} = 0$. In degree 6, the cone $\sigma(\text{⬡})$ is alternating because any automorphism of $M(\text{⬡} = K_4 \cup \{e\})$ fixes e and $\sigma(K_4)$ is alternating, so $V_6^{(4)} = \mathbb{Q}$.

We now explain the nonzero morphisms. Since $V_\bullet^{(4)}$ is 0 in degree less than 2, the map $P_0^{(4)} \to P_{-1}^{(4)}$ is an isomorphism. We now compute the map $P_6^{(4)} \to P_5^{(4)}$. We have that $P_6^{(4)}$ is generated by one cone $\sigma(\boxtimes = K_4 \cup \{e\})$. Its faces are the cones obtained from $K_4 \cup \{e\}$ by deleting one edge. Only deleting the edge e yields a graph which gives an alternating perfect cone, and this cone generates $P_5^{(4)}$. So, this map is an isomorphism.   □

Theorem 6.6 *The top-weight cohomology* $\mathrm{Gr}_{20}^W H^i(\mathcal{A}_4; \mathbb{Q})$ *of* $\mathcal{A}_4$ *is 0 for all* i.

Proof The top-weight cohomology of $\mathcal{A}_4$ is given by the homology of the chain complex $P_\bullet^{(4)}$ by Theorem 3.1. This chain complex has no homology.   □

In fact, our explicit description of $P_\bullet^{(4)}$ shows that $P_\bullet^{(4)} = I_\bullet^{(4)}$, since every nonzero generator is either of rank < 4 or has a coloop. The acyclicity of $P^{(4)}$ is then consistent with Theorem 5.15.

Remark 6.7 Theorem 6.6 can be deduced from the results in [31]. In particular, the weight 0 compactly supported cohomology of $\mathcal{A}_4$ is encoded in the last two columns of [31, Table 1], which describes the first page of a spectral sequence converging to the cohomology of the second Voronoi compactification $\overline{\mathcal{A}}_4^{\mathrm{Vor}}$ of $\mathcal{A}_4$. Here, these two columns contain the compactly supported cohomology of two strata of $\overline{\mathcal{A}}_4^{\mathrm{Vor}}$ whose union is exactly $\mathcal{A}_4$: the fifth column corresponds to the Torelli locus, while the sixth column corresponds to its complement in $\mathcal{A}_4$. By Poincaré duality (10), as described in Section 7, if $\mathcal{A}_4$ had top-weight cohomology it would also have compactly supported cohomology in weight 0. However, even though there are some undetermined entries in the sixth column of the aforementioned table, a close look at the table shows that the weight 0 part must vanish. Indeed, there are no weight 0 classes in the table northwest of the undetermined entries, so any weight 0 classes in the sixth column would persist in the E_∞ page of the spectral sequence and yield weight 0 classes of $\overline{\mathcal{A}}_4^{\mathrm{Vor}}$. But this is impossible as $\overline{\mathcal{A}}_4^{\mathrm{Vor}}$ is a smooth compactification of $\mathcal{A}_4$.

6.3 The complex $P_\bullet^{(5)}$

By using the short exact sequence given in Theorem 4.13, we now compute the complex $P_\bullet^{(5)}$. From this we compute the top-weight cohomology of $\mathcal{A}_5$.

Proposition 6.8 *The chain complex* $P_\bullet^{(5)}$ *is*

$$
\begin{array}{cccccccc}
P_{14}^{(5)} & P_{13}^{(5)} & P_{12}^{(5)} & P_{11}^{(5)} & P_{10}^{(5)} & P_9^{(5)} & P_8^{(5)} & P_7^{(5)}
\end{array}
$$

$$
0 \longrightarrow \mathbb{Q}^3 \xrightarrow{\partial_{14}} \mathbb{Q}^2 \longrightarrow 0 \longrightarrow \mathbb{Q} \xrightarrow{\partial_{11}} \mathbb{Q}^6 \xrightarrow{\partial_{10}} \mathbb{Q}^7 \xrightarrow{\partial_9} \mathbb{Q} \longrightarrow 0
$$

$$
\begin{array}{cccccccc}
P_6^{(5)} & P_5^{(5)} & P_4^{(5)} & P_3^{(5)} & P_2^{(5)} & P_1^{(5)} & P_0^{(5)} & P_{-1}^{(5)}
\end{array}
$$

$$
\longrightarrow \mathbb{Q} \xrightarrow{\sim} \mathbb{Q} \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \mathbb{Q} \xrightarrow{\sim} \mathbb{Q} \longrightarrow 0.
$$

Proof By Theorem 4.13, we have in any degree ℓ that $\dim P_\ell^{(5)} = \dim P_\ell^{(4)} + \dim V_\ell^{(5)}$. We have already computed $P_\bullet^{(4)}$, so we now study $V_\bullet^{(5)}$, which was computed in [22]. Recall from Section 4.1 that Γ_n denotes the set of representatives of alternating perfect cones of dimension $n + 1$. In [22, Table 1], the cardinality of Γ_n, is given by

n	4	5	6	7	8	9	10	11	12	13	14
$|\Gamma_n|$	0	0	0	0	1	7	6	1	0	2	3

In [22, Section 6.2], there is an explicit description of the differential maps.

Since $V_\bullet^{(5)}$ is supported in degrees > 7, while $P_\bullet^{(4)}$ is supported in degrees < 7, the differential maps $P_j^{(5)} \to P_{j-1}^{(5)}$ for $j < 7$ are inherited from $P_\bullet^{(4)}$, and likewise the differential maps $P_j^{(5)} \to P_{j-1}^{(5)}$ for $j > 7$ are inherited from $V_\bullet^{(5)}$. $\qquad\square$

Theorem 6.9 *The top-weight cohomology of $\mathcal{A}_5$ is*

$$\mathrm{Gr}_{30}^W H^i(\mathcal{A}_5; \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } i = 15 \text{ or } 20, \\ 0 & \text{else.} \end{cases}$$

Proof By Proposition 6.8 and [22, Theorem 4.3] we have that $H_9(P_\bullet^{(5)}) = \mathbb{Q}$ and $H_{14}(P_\bullet^{(5)}) = \mathbb{Q}$. Then by Theorem 3.1, we obtain the desired result. $\qquad\square$

Remark 6.10 Grushevsky asks if $\mathcal{A}_g$ ever has nonzero odd cohomology [24, Open Problem 7]. Theorem 6.9 confirms that $\mathcal{A}_5$ does in degree 15. Furthermore, we will see in Theorem 6.12 that Grushevsky's question is also answered affirmatively for $\mathcal{A}_7$, where

$$\dim \mathrm{Gr}_{56}^W H^{33}(\mathcal{A}_7; \mathbb{Q}) = \dim \mathrm{Gr}_{56}^W H^{37}(\mathcal{A}_7; \mathbb{Q}) = 1.$$

6.4 The top-weight cohomology of $\mathcal{A}_6$ and $\mathcal{A}_7$

Elbaz-Vincent, Gangl and Soulé in [22, Theorem 4.3][2] computed the homology of the Voronoi complex $V_\bullet^{(g)}$ for $g = 5$, 6 and 7. Combining this, together with Proposition 6.8, we are able to compute the top-weight cohomology of $\mathcal{A}_6$ and $\mathcal{A}_7$.

Theorem 6.11 *The top-weight cohomology of $\mathcal{A}_6$ is*

$$\mathrm{Gr}_{42}^W H^i(\mathcal{A}_6; \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } i = 30, \\ 0 & \text{else.} \end{cases}$$

Proof By Proposition 4.4, we need to show that $H_{11}(P_\bullet^{(6)}) \cong \mathbb{Q}$ and $H_i(P_\bullet^{(6)}) = 0$ for $i \neq 11$. Consider the long exact sequence in homology arising from the short exact sequence of chain complexes given in

---

[2] Elbaz-Vincent, Gangl and Soulé define the Voronoi complex as a complex of free $\mathbb{Z}$–modules, and in [22, Theorem 4.3] they compute the integral homology of this complex. Our definition of the Voronoi complex $V_\bullet^{(g)}$ is a complex of $\mathbb{Q}$–vector spaces, but this causes no problems as we are only interested in the rational homology of $V_\bullet^{(g)}$.

i	$H_i(P_\bullet^{(5)})$	$H_i(P_\bullet^{(6)})$	$H_i(V_\bullet^{(6)})$
≥ 16	0	0	0
15	0	0	$\mathbb{Q}$
14	$\mathbb{Q}$	0	0
13	0	0	0
12	0	0	0
11	0	$\mathbb{Q}$	$\mathbb{Q}$
10	0	0	$\mathbb{Q}$
9	$\mathbb{Q}$	0	0
≤ 8	0	0	0

Table 2: The long exact sequence in homology for $g = 6$.

Theorem 4.13. Combining this with the computation of the homology of $V_\bullet^{(6)}$ [22, Theorem 4.3] and the homology of $P_\bullet^{(5)}$ given in Proposition 6.8, our computation of $H_k(P_\bullet^{(6)})$ reduces to the four cases in Table 2.

- **Case 1** $(i \leq 8, i = 12, 13, i \geq 16)$ For these values of i, both $H_i(P_\bullet^{(5)})$ and $H_i(V_\bullet^{(6)})$ are equal to zero, so $H_i(P_\bullet^{(6)}) = 0$.

- **Case 2** $(i = 14, 15)$ The long exact sequence in homology gives the exact sequence

$$0 \to H_{15}(P_\bullet^{(6)}) \to \mathbb{Q} \xrightarrow{\delta_{15}^6} \mathbb{Q} \to H_{14}(P_\bullet^{(6)}) \to 0.$$

Exactness implies that the connecting homomorphism δ_{15}^6 is either an isomorphism or the zero map. By [22, Theorem 6.1] we know that inflating the cones in $V_\bullet^{(5)}$ gives an isomorphism of chain complexes $V_\bullet^{(6)} \cong V_\bullet^{(5)}[1] \oplus F_\bullet$ for some complex $F_\bullet$. Combining this with [22, Theorem 4.3] shows the nontrivial homology class in $H_{15}(V_\bullet^{(6)})$ is the inflation of a nontrivial homology class in $H_{14}(V_\bullet^{(5)})$. By Proposition 6.8, the nontrivial homology class in $H_{14}(P_\bullet^{(5)})$ is the nontrivial homology class in $H_{14}(V_\bullet^{(5)})$, so $H_{15}(V_\bullet^{(6)})$ is generated by the inflation of the nontrivial class $H_{14}(P_\bullet^{(5)})$. By the proof of the acyclicity of the inflation complex $I_\bullet^{(g)}$ (Theorem 5.15), this implies the connecting map δ_{15}^6 is an isomorphism. The exact sequence above then implies that $H_k(P_\bullet^{(6)}) = 0$ for both $k = 14$ and $k = 15$.

- **Case 3** $(i = 11)$ $H_{11}(P_\bullet^{(5)})$ and $H_{10}(P_\bullet^{(5)})$ vanish, so the long exact sequence in homology gives

$$0 \to H_{11}(P_\bullet^{(6)}) \to \mathbb{Q} \to 0.$$

This exactness implies that $H_{11}(P_\bullet^{(6)})$ is isomorphic to $\mathbb{Q}$.

- **Case 4** $(i = 9, 10)$ By considering the long exact sequence in homology in the range $i = 10$ to $i = 9$ we have the exact sequence

$$0 \to H_{10}(P_\bullet^{(6)}) \to \mathbb{Q} \xrightarrow{\delta_{10}^6} \mathbb{Q} \to H_9(P_\bullet^{(6)}) \to 0.$$

An analysis similar to that in Case 2 shows that connecting map δ_{11}^6 is an isomorphism, implying by exactness that $H_i(P_\bullet^{(6)}) = 0$ for both $i = 9$ and $i = 10$. □

i	$H_i(P_\bullet^{(6)})$	$H_i(P_\bullet^{(7)})$	$H_i(V_\bullet^{(7)})$
≥ 28	0	0	0
27	0	$\mathbb{Q}$	$\mathbb{Q}$
26	0	0	0
25	0	0	0
24	0	0	0
23	0	0	0
22	0	$\mathbb{Q}$	$\mathbb{Q}$
21	0	0	0
20	0	0	0
19	0	0	0
18	0	$\mathbb{Q}$	$\mathbb{Q}$
17	0	0	0
16	0	0	0
15	0	0	0
14	0	0	0
13	0	$\mathbb{Q}$	$\mathbb{Q}$
12	0	0	$\mathbb{Q}$
11	$\mathbb{Q}$	0	0
10	0	0	0
9	0	0	0
≤ 8	0	0	0

Table 3: The long exact sequence in homology for $g = 7$.

We now compute the top-weight rational cohomology of $\mathcal{A}_7$.

Theorem 6.12 *The top-weight cohomology of $\mathcal{A}_7$ is*

$$\mathrm{Gr}_{56}^W H^i(\mathcal{A}_7; \mathbb{Q}) = \begin{cases} \mathbb{Q} & \text{if } i = 28, 33, 37, 42, \\ 0 & \text{else.} \end{cases}$$

Proof We compute the homology of $P_\bullet^{(7)}$ in a similar fashion to the proof of Theorem 6.11, by considering the long exact sequence in homology arising from the short exact sequence of chain complexes given in Theorem 4.13. Table 3 records the homology of $P_\bullet^{(6)}$ and $V_\bullet^{(7)}$, which are given in Table 2 and [22, Theorem 4.3], respectively.

Both Case 1 ($i \neq 11, 12, 13, 18, 22, 27$) and Case 2 ($i = 13, 18, 22, 27$) follow from the exactness of the long exact sequence on homology in a manner analogous to Cases 1 and 3 in the proof of Theorem 6.11.

For Case 3 ($i = 11, 12$), the long exact sequence in homology gives the exact sequence

$$0 \to H_{12}(P_\bullet^{(7)}) \to \mathbb{Q} \xrightarrow{\delta_{12}^7} \mathbb{Q} \to H_{11}(P_\bullet^{(7)}) \to 0.$$

Now δ_{12}^7 is either an isomorphism or it is the zero map. As discussed in [22, Section 6.3], the nontrivial homology class in $H_{12}(V_\bullet^{(7)})$ is the inflation of a nontrivial homology class in $H_{11}(V_\bullet^{(6)})$. However, since

by the proof of Theorem 6.11, the nontrivial homology class in $H_{11}(P_\bullet^{(6)})$ is the nontrivial homology class in $H_{11}(V_\bullet^{(6)})$, this implies that $H_{12}(V_\bullet^{(7)})$ is generated by the inflation of the nontrivial class $H_{11}(P_\bullet^{(6)})$. By the proof of the acyclicity of the inflation complex $I^{(g)}$ (Theorem 6.11), this implies the connecting map δ_{12}^7 is an isomorphism. The exact sequence above then implies $H_k(P_\bullet^{(7)}) = 0$ for $i = 11$ and $i = 12$.   □

Theorem A now follows directly from Theorems 6.6, 6.9, 6.11 and 6.12. As a corollary of this we are able to deduce the top-weight Euler characteristic of $\mathcal{A}_g$ for $2 \leq g \leq 7$.

Corollary 6.13 *The top-weight Euler characteristic of $\mathcal{A}_g$ for $2 \leq g \leq 7$ is*

$$\chi^{\text{top}}(\mathcal{A}_g) = \begin{cases} 1 & \text{if } g = 3, 6 \\ 0 & \text{if } g = 2, 4, 5, 7. \end{cases}$$

Remark 6.14 One can also deduce the top-weight Euler characteristic of $\mathcal{A}_g$ for $5 \leq g \leq 7$ directly from the numbers listed in [22, Figures 1 and 2]. It would be interesting to know whether a closed formula for the top-weight Euler characteristic of $\mathcal{A}_g$ exists in general.

Remark 6.15 We have established

$$(7) \qquad\qquad\qquad \text{Gr}_{(g+1)g}^W H^{g(g-1)}(\mathcal{A}_g; \mathbb{Q}) \neq 0$$

for $g = 3, 5, 6$ and 7 ($g = 3$ also follows from [27]). We ask whether (7) holds for all $g \geq 5$. Equivalently, the question is whether $H_{2g-1}(P^{(g)}) \neq 0$ for all $g \geq 5$. The connection to the stable cohomology of the Satake compactification, as summarized in Table 4, gives evidence for this question, as explained in Section 7; see Question 7.1. We also note the possible relationship with the main theorems of [12] on the rational cohomology of $\mathcal{M}_g$, which use the fact that $H_{2g-1}(G^{(g)}) \neq 0$ for $g = 3$ and $g \geq 5$; see [9], [46] and [12, Theorem 2.7]. We leave this interesting investigation as an open question.

6.5   Results for $g \geq 8$

While full calculations for the top-weight cohomology of $\mathcal{A}_g$ in the range $g \geq 8$ are beyond the scope of current computations, we can nevertheless use our previous computation of the top-weight cohomology of $\mathcal{A}_7$ together with a vanishing result of [21] to show that the top-weight cohomology of $\mathcal{A}_8, \mathcal{A}_9$ and $\mathcal{A}_{10}$ vanishes in a certain range slightly larger than what is given by the virtual cohomological dimension.

Theorem 6.16 *The top-weight rational cohomology of $\mathcal{A}_8, \mathcal{A}_9$ and $\mathcal{A}_{10}$ vanishes in the ranges*

$$\begin{aligned}
\text{Gr}_{72}^W H^i(\mathcal{A}_8; \mathbb{Q}) &= 0 \quad \text{for } i \geq 60, \\
\text{Gr}_{90}^W H^i(\mathcal{A}_9; \mathbb{Q}) &= 0 \quad \text{for } i \geq 79, \\
\text{Gr}_{110}^W H^i(\mathcal{A}_{10}; \mathbb{Q}) &= 0 \quad \text{for } i \geq 99.
\end{aligned}$$

(8)

Proof By Theorem 4.5 of [21] for $g = 8, 9$ and 10 the homology $H_i(V_\bullet^{(g)}) = 0$ for $i \leq 11$, and further, $H_{12}(V_\bullet^{(8)}) = 0$. Considering the long exact sequence in homology

$$\cdots \to H_{i+1}(V_\bullet^{(g)}) \xrightarrow{\delta} H_i(P_\bullet^{(g-1)}) \to H_i(P_\bullet^{(g)}) \to H_i(V_\bullet^{(g)}) \to \cdots$$

coming from the short exact sequence of chain complexes given in Theorem 4.13, we see that this vanishing implies that

$$H_i(P_\bullet^{(7)}) \cong H_i(P_\bullet^{(8)}) \cong H_i(P_\bullet^{(9)}) \cong H_i(P_\bullet^{(10)}) \quad \text{for } i \leq 10$$

and $H_{11}(P_\bullet^{(7)}) \cong H_{11}(P_\bullet^{(8)})$. By our computation of the homology of $P_\bullet^{(7)}$ in the proof of Theorem 6.12 we know that $H_i(P_\bullet^{(7)}) = 0$ for all $i \leq 12$, implying that for $g = 8, 9$ and 10, the homology $H_i(P_\bullet^{(g)}) = 0$ for $i \leq 10$, and further, $H_{11}(P_\bullet^{(8)}) = 0$. The result now follows from Proposition 4.4. □

Remark 6.17 These vanishing bounds for $g = 8, 9, 10$ are slightly larger than the bounds provided by Corollary 5.16, equivalently, the fact that vcd $\mathcal{A}_g = g^2$ (see Remark 5.19), which imply that

$$\operatorname{Gr}_{72}^W H^i(\mathcal{A}_8; \mathbb{Q}) = 0 \quad \text{for } i \geq 65,$$
$$\operatorname{Gr}_{90}^W H^i(\mathcal{A}_9; \mathbb{Q}) = 0 \quad \text{for } i \geq 82,$$
$$\operatorname{Gr}_{110}^W H^i(\mathcal{A}_{10}; \mathbb{Q}) = 0 \quad \text{for } i \geq 101.$$

The result for $g = 10$, however, is subsumed by the more general fact that the top-weight cohomology of $\mathcal{A}_g$ vanishes in degrees 0 and 1 below the vcd, as we shall note in Section 7 below.

7 Relationship with the stable cohomology of $\mathcal{A}_g^{\mathrm{Sat}}$

Our results on the existence of certain top-weight cohomology classes of $\mathcal{A}_g$ can be related to results of Chen and Looijenga [15] and Charney and Lee [14] which predict that, as g grows, there should be infinitely many of these classes. This connection was brought to our attention by O Tommasi, and we thank her for explaining her ideas to us in detail.

Recall that $\mathcal{A}_g$ admits a compactification $\mathcal{A}_g^{\mathrm{Sat}}$, called the Satake or Baily–Borel compactification, first constructed as a projective variety by Baily and Borel in [5]. This compactification can be seen as a minimal compactification in the sense that it admits a morphism from all toroidal compactifications of $\mathcal{A}_g$. The reader interested in learning more about the vast literature on $\mathcal{A}_g$ and its compactifications can look at the very nice surveys [24; 32]. There are natural maps $\mathcal{A}_g^{\mathrm{Sat}} \to \mathcal{A}_{g+1}^{\mathrm{Sat}}$, and the groups $H^k(\mathcal{A}_g^{\mathrm{Sat}}; \mathbb{Q})$ stabilize for $k < g$; see [14]. Moreover, as Charney and Lee prove, the stable cohomology ring $H^\bullet(\mathcal{A}_\infty^{\mathrm{Sat}}; \mathbb{Q})$ of the Satake compactifications is freely generated by the classes λ_i for i odd, and the classes y_{4j+2} for $j = 1, 2, 3, \ldots$, where y_{4j+2} is in degree $4j + 2$. Here, the λ–classes extend the ith Chern class of the Hodge bundle on $\mathcal{A}_g$; in particular they are algebraic, and hence never have weight 0. But the classes y_j have weight 0, as proven recently by Chen and Looijenga [15]. This result is very important in the discussion that follows.

Recall also that $\mathcal{A}_g^{\mathrm{Sat}}$ admits a stratification by locally closed substacks

$$\mathcal{A}_g^{\mathrm{Sat}} = \mathcal{A}_g \sqcup \mathcal{A}_{g-1} \sqcup \cdots \sqcup \mathcal{A}_0.$$

Thus the spectral sequence on compactly supported cohomology associated to this stratification is

(9) $$E_1^{p,q} = H_c^{p+q}(\mathcal{A}_p; \mathbb{Q}) \Rightarrow H^{p+q}(\mathcal{A}_g^{\mathrm{Sat}}; \mathbb{Q}),$$

where $p = 0, \ldots, g$. This spectral sequence may be interpreted in the category of mixed Hodge structures. Passing to the weight 0 subspace, we see that the existence of the products of the y_j classes in the stable cohomology ring of the Satake compactification implies the existence of infinitely many cohomology classes in $\mathrm{Gr}_0^W H_c^j(\mathcal{A}_g; \mathbb{Q})$ for all g, and hence by the perfect pairing

(10) $$\mathrm{Gr}_0^W H_c^j(\mathcal{A}_g; \mathbb{Q}) \times \mathrm{Gr}_{(g+1)g}^W H^{(g+1)g-j}(\mathcal{A}_g; \mathbb{Q}) \to \mathbb{Q}$$

provided by Poincaré duality, infinitely many classes $\mathrm{Gr}_{(g+1)g}^W H^*(\mathcal{A}_g; \mathbb{Q})$ in top weight.

With Poincaré duality applied, all of the known results on the top-weight cohomology of $\mathcal{A}_g$, including our Theorems 6.9, 6.11 and 6.12, can thus be summarized in Table 4, which shows the weight 0 part of the E_1 page of the spectral sequence (9).

Implicit in Table 4 is the fact that all terms below the p–axis are zero. This follows from the fact that $\mathrm{vcd}(\mathcal{A}_g) = \mathrm{vcd}(\mathrm{Sp}(2g, \mathbb{Z})) = g^2$, or, just as well, from the fact that $\mathrm{vcd}(\mathrm{GL}_g(\mathbb{Z})) = \binom{g}{2}$; see [7]. In fact, the vanishing below the p–axis as well as in the rows $q = 0, 1$ and 2, apart from $(p, q) = (0, 0)$, can be deduced from the fact that the cohomology of $\mathrm{GL}_g(\mathbb{Z})$ with coefficients in $\widetilde{\mathbb{Q}}$ vanishes in degrees 0, 1 and 2 below the vcd. Indeed, we have, for all k,

$$H^{\binom{g}{2}-k}(\mathrm{GL}_g(\mathbb{Z}); \widetilde{\mathbb{Q}}) \cong H_k(\mathrm{GL}_g(\mathbb{Z}); \mathrm{St} \otimes \mathbb{Q}) \cong H_{k+g-1}(V^{(g)}),$$

where St denotes the Steinberg module [43]; these are all zero when $g > 1$ for $k = 0$ (see [35]), $k = 1$ (see [17]), and $k = 2$ (see [10]). Then Theorem 4.13 implies that also $H_{k+g-1}(P^{(g)}) = 0$ for $k = 0, 1$ and 2 so also $\mathrm{Gr}_{g^2+g}^W H^{g^2-k}(\mathcal{A}_g; \mathbb{Q}) = (\mathrm{Gr}_0^W H_c^{g+k}(\mathcal{A}_g; \mathbb{Q}))^\vee = 0$ for $g > 0$ and $k \leq 2$ by Proposition 4.4.

As explained to us by Tommasi, the classes in Theorems 6.9, 6.11 and 6.12, as well as the already-known class in $\mathrm{Gr}_{12}^W H^6(\mathcal{A}_3; \mathbb{Q})$ from [27] give natural candidates for classes in $\mathrm{Gr}_0^W H_c^{p+q}(\mathcal{A}_p; \mathbb{Q})$ that produce the classes y_{4j+2} in the spectral sequence (9), in the sense that they persist in the E_∞ page in the Gysin spectral sequence for g sufficiently large. Indeed, looking at the $p = q$ diagonal on the E_1 page of the spectral sequence in Table 4, we are led to ask:

Question 7.1   (i)   Is $\mathrm{Gr}_0^W H_c^{2g}(\mathcal{A}_g; \mathbb{Q}) \neq 0$ for $g = 3$ and all $g \geq 5$?

  (ii)   Do these cohomology classes produce the stable cohomology classes in $\mathrm{Gr}_0^W H^\bullet(\mathcal{A}_\infty^{\mathrm{Sat}}; \mathbb{Q})$?

  (iii)   Is $\mathrm{Gr}_0^W H_c^k(\mathcal{A}_g; \mathbb{Q}) = 0$ for $k < 2g$?

As discussed in the introduction, an affirmative answer to the third question in the range $k < 2g - 1$ would be implied by [16, Conjecture 2]. Our Theorems 6.2, 6.9, 6.11 and 6.12 verify the first and third questions

	0	1	2	3	4	5	6	7	8	9	10	
21	0	0	0	0	0	0	0	$\mathbb{Q}$				
20	0	0	0	0	0	0	0	0				
19	0	0	0	0	0	0	0	0				
18	0	0	0	0	0	0	0	0				
17	0	0	0	0	0	0	0	0				
16	0	0	0	0	0	0	0	$\mathbb{Q}$				
15	0	0	0	0	0	0	0	0				
14	0	0	0	0	0	0	0	0				
13	0	0	0	0	0	0	0	0				
12	0	0	0	0	0	0	0	$\mathbb{Q}$				
11	0	0	0	0	0	0	0	0				
10	0	0	0	0	0	$\mathbb{Q}$	0	0				
9	0	0	0	0	0	0	0	0				
8	0	0	0	0	0	0	0	0				
7	0	0	0	0	0	0	0	$\mathbb{Q}$				
6	0	0	0	0	0	0	$\mathbb{Q}$	0				
5	0	0	0	0	0	$\mathbb{Q}$	0	0				
4	0	0	0	0	0	0	0	0	0			
3	0	0	0	$\mathbb{Q}$	0	0	0	0	0			
2	0	0	0	0	0	0	0	0	0	0		
1	0	0	0	0	0	0	0	0	0	0	0	$\cdots$
0	$\mathbb{Q}$	0	0	0	0	0	0	0	0	0	0	$\cdots$

Table 4: The page $E_1^{p,q} = \mathrm{Gr}_0^W H_c^{p+q}(\mathcal{A}_p; \mathbb{Q}) \Rightarrow \mathrm{Gr}_0^W H^{p+q}(\mathcal{A}_g^{\mathrm{Sat}}; \mathbb{Q})$ of the Gysin spectral sequence, for g sufficiently large. The blank entries for $p \geq 8$ are currently unknown.

for $g \leq 7$. They also verify the second question for $g = 3$ and for $g = 5$. Indeed, $\mathrm{Gr}_0^W H_c^6(\mathcal{A}_3; \mathbb{Q})$ and $\mathrm{Gr}_0^W H_c^{10}(\mathcal{A}_5; \mathbb{Q})$ are the only nonzero terms in the antidiagonals $p+q = 6$ and $p+q = 10$, respectively; so they produce the classes $y_6 \in \mathrm{Gr}_0^W H^6(\mathcal{A}_\infty^{\mathrm{Sat}}; \mathbb{Q})$ and $y_{10} \in \mathrm{Gr}_0^W H^{10}(\mathcal{A}_\infty^{\mathrm{Sat}}; \mathbb{Q})$, respectively. It is natural to guess that the other terms in Table 4 similarly produce products of the y_j: for example, that $\mathrm{Gr}_0^W H_c^{12}(\mathcal{A}_6; \mathbb{Q})$ produces y_6^2, and that $\mathrm{Gr}_0^W H_c^{14}(\mathcal{A}_7; \mathbb{Q})$ produces y_{14}, and so on.

Finally, Tommasi also remarks that the odd-degree classes in weight 0 compactly supported cohomology of $\mathcal{A}_g$ detected so far, namely

$$\mathrm{Gr}_0^W H_c^{15}(\mathcal{A}_5; \mathbb{Q}), \quad \mathrm{Gr}_0^W H_c^{19}(\mathcal{A}_7; \mathbb{Q}) \quad \text{and} \quad \mathrm{Gr}_0^W H_c^{23}(\mathcal{A}_7; \mathbb{Q}),$$

must of course be killed by a differential on some page of the spectral sequence, since $\mathcal{A}_g^{\mathrm{Sat}}$ has no weight 0 stable cohomology in odd degrees. This implies the existence of some even-degree classes in $\mathrm{Gr}_0^W H_c^\bullet(\mathcal{A}_g; \mathbb{Q})$ which kill the odd-degree classes and which are not related by this spectral sequence to the products of the y_j. It would be very interesting to explicitly identify such classes.

References

[1] **D Abramovich**, **L Caporaso**, **S Payne**, *The tropicalization of the moduli space of curves*, Ann. Sci. Éc. Norm. Supér. 48 (2015) 765–809 MR Zbl

[2] **V Alexeev**, **A Brunyate**, *Extending the Torelli map to toroidal compactifications of Siegel space*, Invent. Math. 188 (2012) 175–196 MR Zbl

[3] **D Allcock**, **D Corey**, **S Payne**, *Tropical moduli spaces as symmetric Δ–complexes*, Bull. Lond. Math. Soc. 54 (2022) 193–205 MR Zbl

[4] **A Ash**, **D Mumford**, **M Rapoport**, **Y Tai**, *Smooth compactification of locally symmetric varieties*, Lie Groups: History, Frontiers and Applications 4, Math. Sci. Press, Brookline, MA (1975) MR Zbl

[5] **W L Baily, Jr**, **A Borel**, *Compactification of arithmetic quotients of bounded symmetric domains*, Ann. of Math. 84 (1966) 442–528 MR Zbl

[6] **A Borel**, *Stable real cohomology of arithmetic groups*, Ann. Sci. École Norm. Sup. 7 (1974) 235–272 MR Zbl

[7] **A Borel**, **J-P Serre**, *Corners and arithmetic groups*, Comment. Math. Helv. 48 (1973) 436–491 MR Zbl

[8] **S Brannetti**, **M Melo**, **F Viviani**, *On the tropical Torelli map*, Adv. Math. 226 (2011) 2546–2586 MR Zbl

[9] **F Brown**, *Mixed Tate motives over $\mathbb{Z}$*, Ann. of Math. 175 (2012) 949–976 MR Zbl

[10] **B Brück**, **J Miller**, **P Patzt**, **R J Sroka**, **J C H Wilson**, *On the codimension-two cohomology of* $SL_n(\mathbb{Z})$, preprint (2022) arXiv 2204.11967

[11] **M Chan**, *Combinatorics of the tropical Torelli map*, Algebra Number Theory 6 (2012) 1133–1169 MR Zbl

[12] **M Chan**, **S Galatius**, **S Payne**, *Tropical curves, graph complexes, and top weight cohomology of* $\mathcal{M}_g$, J. Amer. Math. Soc. 34 (2021) 565–594 MR Zbl

[13] **M Chan**, **M Melo**, **F Viviani**, *Tropical Teichmüller and Siegel spaces*, from "Algebraic and combinatorial aspects of tropical geometry" (E Brugallé, M A Cueto, A Dickenstein, E-M Feichtner, I Itenberg, editors), Contemp. Math. 589, Amer. Math. Soc., Providence, RI (2013) 45–85 MR Zbl

[14] **R Charney**, **R Lee**, *Cohomology of the Satake compactification*, Topology 22 (1983) 389–423 MR Zbl

[15] **J Chen**, **E Looijenga**, *The stable cohomology of the Satake compactification of* $\mathcal{A}_g$, Geom. Topol. 21 (2017) 2231–2241 MR Zbl

[16] **T Church**, **B Farb**, **A Putman**, *A stability conjecture for the unstable cohomology of* $SL_n\mathbb{Z}$, *mapping class groups, and* $\mathrm{Aut}(F_n)$, from "Algebraic topology: applications and new directions" (U Tillmann, S r Galatius, D Sinha, editors), Contemp. Math. 620, Amer. Math. Soc., Providence, RI (2014) 55–70 MR Zbl

[17] **T Church**, **A Putman**, *The codimension-one cohomology of* $SL_n\mathbb{Z}$, Geom. Topol. 21 (2017) 999–1032 MR Zbl

[18] **D A Cox**, **J B Little**, **H K Schenck**, *Toric varieties*, Graduate Studies in Math. 124, Amer. Math. Soc., Providence, RI (2011) MR Zbl

[19] **P Deligne**, *Théorie de Hodge, II*, Inst. Hautes Études Sci. Publ. Math. 40 (1971) 5–57 MR Zbl

[20] **P Deligne**, *Théorie de Hodge, III*, Inst. Hautes Études Sci. Publ. Math. 44 (1974) 5–77 MR Zbl

[21] **M Dutour Sikirić**, **P Elbaz-Vincent**, **A Kupers**, **J Martinet**, *Voronoi complexes in higher dimensions, cohomology of* $GL_N(Z)$ *for* $N \geq 8$ *and the triviality of* $K_8(Z)$, preprint (2019) arXiv 1910.11598

[22]   **P Elbaz-Vincent**, **H Gangl**, **C Soulé**, *Perfect forms, K–theory and the cohomology of modular groups*, Adv. Math. 245 (2013) 587–624   MR Zbl

[23]   **G Faltings**, **C-L Chai**, *Degeneration of abelian varieties*, Ergebnisse der Math. (3) 22, Springer (1990) MR Zbl

[24]   **S Grushevsky**, *Geometry of $\mathcal{A}_g$ and its compactifications*, from "Algebraic geometry, I" (D Abramovich, A Bertram, L Katzarkov, R Pandharipande, M Thaddeus, editors), Proc. Sympos. Pure Math. 80.1, Amer. Math. Soc., Providence, RI (2009) 193–234   MR Zbl

[25]   **S Grushevsky**, **K Hulek**, **O Tommasi**, *Stable Betti numbers of (partial) toroidal compactifications of the moduli space of abelian varieties*, from "Geometry and physics, II" (J E Andersen, A Dancer, O García-Prada, editors), Oxford Univ. Press (2018) 581–609   MR Zbl

[26]   **S Grushevsky**, **K Hulek**, **O Tommasi**, *Stable cohomology of the perfect cone toroidal compactification of $\mathcal{A}_g$*, J. Reine Angew. Math. 741 (2018) 211–254   MR Zbl

[27]   **R Hain**, *The rational cohomology ring of the moduli space of abelian 3–folds*, Math. Res. Lett. 9 (2002) 473–491   MR Zbl

[28]   **F Harary**, **M J Piff**, **D J A Welsh**, *On the automorphism group of a matroid*, Discrete Math. 2 (1972) 163–171   MR Zbl

[29]   **G Harder**, *A Gauss–Bonnet formula for discrete arithmetically defined groups*, Ann. Sci. École Norm. Sup. 4 (1971) 409–455   MR Zbl

[30]   **A Hatcher**, *Algebraic topology*, Cambridge Univ. Press (2002)   MR Zbl

[31]   **K Hulek**, **O Tommasi**, *Cohomology of the second Voronoi compactification of $\mathcal{A}_4$*, Doc. Math. 17 (2012) 195–244   MR Zbl

[32]   **K Hulek**, **O Tommasi**, *The topology of $\mathcal{A}_g$ and its compactifications*, from "Geometry of moduli" (J A Christophersen, K Ranestad, editors), Abel Symp. 14, Springer (2018) 135–193   MR Zbl

[33]   **J Igusa**, *On Siegel modular forms of genus two*, Amer. J. Math. 84 (1962) 175–200   MR

[34]   **D-O Jaquet-Chiffelle**, *Énumération complète des classes de formes parfaites en dimension 7*, Ann. Inst. Fourier (Grenoble) 43 (1993) 21–55   MR Zbl

[35]   **R Lee**, **R H Szczarba**, *On the homology and cohomology of congruence subgroups*, Invent. Math. 33 (1976) 15–53   MR Zbl

[36]   **R Lee**, **R H Szczarba**, *On the torsion in $K_4(\mathbb{Z})$ and $K_5(\mathbb{Z})$*, Duke Math. J. 45 (1978) 101–129   MR Zbl

[37]   **M Melo**, **F Viviani**, *Comparing perfect and 2nd Voronoi decompositions: the matroidal locus*, Math. Ann. 354 (2012) 1521–1554   MR Zbl

[38]   **Y Namikawa**, *Toroidal compactification of Siegel spaces*, Lecture Notes in Math. 812, Springer (1980)   MR Zbl

[39]   **Y Odaka**, *Tropical geometric compactification of moduli, II: $\mathcal{A}_g$ case and holomorphic limits*, Int. Math. Res. Not. 2019 (2019) 6614–6660   MR Zbl

[40]   **Y Odaka**, **Y Oshima**, *Collapsing K3 surfaces, tropical geometry and moduli compactifications of Satake, Morgan–Shalen type*, MSJ Memoirs 40, Math. Soc. Japan, Tokyo (2021)   MR Zbl

[41]   **J G Oxley**, *Matroid theory*, Oxford Univ. Press (1992)   MR Zbl

[42]   **N I Shepherd-Barron**, *Perfect forms and the moduli space of abelian varieties*, Invent. Math. 163 (2006) 25–45   MR Zbl

[43]　**C Soulé**, *On the 3–torsion in $K_4(\mathbb{Z})$*, Topology 39 (2000) 259–265　MR　Zbl

[44]　**G Voronoi**, *Nouvelles applications des paramètres continus à la théorie des formes quadratiques, I: Sur quelques propriétés des formes quadratiques positives parfaites*, J. Reine Angew. Math. 133 (1908) 97–102　MR　Zbl

[45]　**H Whitney**, *Congruent graphs and the connectivity of graphs*, Amer. J. Math. 54 (1932) 150–168　MR　Zbl

[46]　**T Willwacher**, *M Kontsevich's graph complex and the Grothendieck–Teichmüller Lie algebra*, Invent. Math. 200 (2015) 671–760　MR　Zbl

Department of Mathematics, Brown University
Providence, RI, United States

Department of Mathematics, University of California, Berkeley
Berkeley, CA, United States

Current address:　*Department of Mathematics, Brown University*
Providence, RI, United States

Department of Mathematics, Brown University
Providence, RI, United States

Department of Mathematics and Physics, Università Roma Tre
Rome, Italy

Department of Mathematics, Harvard University
Cambridge, MA, United States

Current address:　*Department of Mathematics, Statistics and Computer Science, University of Illinois Chicago*
Chicago, IL, United States

Department of Mathematics, Tulane University
New Orleans, LA, United States

madeline_brandt@brown.edu,　juliette_bruce1@brown.edu,　melody_chan@brown.edu,
melo@mat.uniroma3.it,　gwynm@uic.edu,　cwolfe@tulane.edu

Algebraic uniqueness of Kähler–Ricci flow limits and optimal degenerations of Fano varieties

JIYUAN HAN

CHI LI

We prove that for any Fano manifold X, the special $\mathbb{R}$–test configuration that minimizes the $\boldsymbol{H}^{\mathrm{NA}}$–functional is unique and has a K–semistable $\mathbb{Q}$–Fano central fiber (W, ξ). Moreover there is a unique K–polystable degeneration of (W, ξ). As an application, we confirm the conjecture of Chen, Sun and Wang about the algebraic uniqueness for Kähler–Ricci flow limits on Fano manifolds, which implies that the Gromov–Hausdorff limit of the flow does not depend on the choice of initial Kähler metrics. The results are achieved by studying algebraic optimal degeneration problems via new functionals for real valuations over $\mathbb{Q}$–Fano varieties, which are analogous to the minimization problem for normalized volumes.

1 Introduction

Let X be a smooth Fano manifold. It is now known that X admits a Kähler–Einstein metric if and only if X is K–polystable; see Berman [5], Chen, Donaldson and Sun [25; 26; 27] and Tian [67; 68]. In this paper, we are interested in the case when X is not K–polystable. If X is strictly K–semistable,

then X admits a unique K–polystable degeneration by Li, Wang and Xu [55]. If X is K–unstable (ie not K–semistable), several kinds of optimal degenerations have been studied which are related to continuity methods or geometric flows in the analytic study of canonical metrics. For example, related to Aubin's continuity method, there is a (not necessarily unique) special degeneration whose associated valuation minimizes the δ invariant; see Blum, Liu and Zhou [16] and Székelyhidi [65]. There is also a unique destabilizing geodesic ray which arises in the study of inverse Monge–Ampère flow (resp. Calabi flow) and whose associated non-Archimedean metric minimizes an L^2–normalized non-Archimedean Ding invariant (resp. L^2–normalized radial Calabi functional); see Donaldson [33], Hisamoto [42] and Xia [77]. In this paper we are interested in optimal degenerations that arise in the study of Hamilton–Tian conjecture about the long time behavior of Kähler–Ricci flows. The latter conjecture states that starting from any Kähler metric $\omega \in c_1(X)$, the normalized Kähler–Ricci converges in the Gromov–Hausdorff sense to a Kähler–Ricci soliton on a $\mathbb{Q}$–Fano variety X_∞. The Hamilton–Tian conjecture has been solved (see Bamler [4], Chen and Wang [29] and Tian and Zhang [70]) and applied to give a proof of the Yau–Tian–Donaldson conjecture in Chen, Sun and Wang [28].

It is known that X_∞ coincides with X if and only if there is already a Kähler–Ricci soliton on X; see Dervan and Székelyhidi [31] and Tian and Zhu [72]. In general, Chen, Sun and Wang [28] proved the following phenomenon. The metric degeneration from X to X_∞ induces a finitely generated filtration $\mathcal{F}$ on $R = \bigoplus_m H^0(X, -mK_X)$, and there is a two-step degeneration:

(i) The filtration $\mathcal{F}$ as an $\mathbb{R}$–test configuration (see Definition 2.8) degenerates X to a normal Fano variety W with a torus $\mathbb{T}$–action generated by a holomorphic vector field ξ. For simplicity, we call this step the semistable degeneration.

(ii) There is a $\mathbb{T}$–equivariant test configuration of (W, ξ) to (X_∞, ξ). We call this step the polystable degeneration.

As explained in Chen, Sun and Wang [28], this picture is a global analogue of the picture in Donaldson and Sun's study [34] of metric tangent cones on Gromov–Hausdorff limits of Fano Kähler–Einstein manifolds. In [34], Donaldson and Sun conjectured that metric tangent cones depend only on the algebraic structure near the singularity. This conjecture has been confirmed in a series of works of the second author with his collaborators (see Li [51], Li and Xu [58; 57] and Li, Wang and Xu [55]), which depends on the study of the minimization problem of a normalized volume functional over the space of valuations centered at the singularity; see Li, Liu and Xu [54] for a survey. Analogous to this conjecture on metric tangent cones, the following conjecture was proposed in [28]:

Conjecture 1.1 *The data $\mathcal{F}$, W and X_∞ depend only on the algebraic structure of X but not on the initial metric for the Kähler–Ricci flow.*

In this paper we will confirm Conjecture 1.1. The idea and method to prove this conjecture are in some sense parallel to the study of minimizing normalized volumes. However, the correct framework for achieving this goal has not been established until now. So the second purpose of this paper is to study

an analogous minimization problem in the global setting, which can be studied for all $\mathbb{Q}$–Fano varieties possibly singular, and prove various results about it.

The functional we want to minimize is called the H^{NA}–functional of $\mathbb{R}$–test configurations.[1] Tian, Zhang, Zhang and Zhu [69, Proposition 5.1] first introduced the H^{NA}–functional for holomorphic vector fields in their study of Kähler–Ricci flow on Fano manifolds. This invariant was generalized to any *special* $\mathbb{R}$–test configuration by Dervan and Székelyhidi [31], who then used the results of Chen and Wang [29], Chen, Sun and Wang [28] and He [40] to prove that the semistable degeneration mentioned above minimizes the H^{NA}–functional among all special $\mathbb{R}$–test configurations; see Remark 2.44. For general test configurations, such an H^{NA}–functional is a nonlinear version of the non-Archimedean Berman–Ding functional, and was first explicitly used by Hisamoto in [43] to reprove Dervan and Székelyhidi's result using pluripotential theory. Note that in this paper, for the convenience of our argument and comparison with the case of the δ invariant (or with the β invariant, see equation (107)), we will use the negative of the sign convention in these previous works.

Conjecture 1.1 follows from two purely algebrogeometric statements for each step of the semistable and polystable degenerations.

Theorem 1.2   *For any $\mathbb{Q}$–Fano variety, the special $\mathbb{R}$–test configuration that minimizes H^{NA} is unique and its central fiber (W, ξ) is K–semistable (Definition 2.49).*

Theorem 1.3   *If (X, ξ) is K–semistable, then there exists a unique K–polystable degeneration.*

Corollary 1.4   *Conjecture 1.1 is true for any smooth Fano manifold. In particular, the Gromov–Hausdorff limit X_∞ for the Kähler–Ricci flow does not depend on the initial metric of the flow.*

To prepare for the proof of such results, we will first carry out an algebraic study of the H^{NA}–functional, which is analogous to the study of the minimization problem for normalized volume or the δ invariant. We will prove a new interesting fact in Theorem 3.5, that the MMP process devised in [56] decreases the H^{NA} invariant of test configurations. This requires us to derive new intersection formulas (see (121)) and derivative formulas for the H^{NA} invariant. The proof of such formulas depends on a fibration technique in the study of equivariant cohomology. This technique is partly motivated by some construction from our previous work [39], although there are key differences which require more concrete calculations; see Remark 3.2.

We will then introduce the following $\widetilde{\beta}$–functional on $\mathrm{Val}(X)$ the space of valuations on X: for any $v \in \mathrm{Val}(X)$ with $A_X(v) < +\infty$, we define

$$(1) \qquad \widetilde{\beta}(v) = A_X(v) + \log\left(\frac{1}{(-K_X)^{\cdot n}} \int_{\mathbb{R}} e^{-\lambda} (-d\,\mathrm{vol}(\mathcal{F}_v^{(\lambda)})) \right).$$

---

[1]We will mostly use the notation of non-Archimedean functionals, as advocated in Boucksom, Hisamoto and Jonsson [19]. However, note that H^{NA} here is not the non-Archimedean entropy functional used in [19]. We will not use the non-Archimedean entropy in this article.

See Section 4 for details. If $A_X(v) = +\infty$, then we define $\widetilde{\beta}(v) = +\infty$. The transition from $\boldsymbol{H}^{\mathrm{NA}}$ to the $\widetilde{\beta}$–functional is similar to the transition from Ding–Tian's generalized Futaki invariant to the β–functional in the literature of K–stability (as first appeared in Li [50], where it was called Θ, and in Fujita [37]). In other words, $\widetilde{\beta}$ is a nonlinear version of β and could also be considered as a global analogue of the normalized volume. Unlike the case of normalized volume functional, the $\widetilde{\beta}$ invariant is not invariant under rescaling of valuations. Indeed, we find the following new phenomenon: when restricted to the ray of multiples of a fixed valuation $v \in \mathrm{Val}(X)$ with $A(v) < +\infty$, it is strictly convex and proper and its derivative at the origin is exactly $\beta(v)$. As a consequence there is a unique minimizer along the ray, which is nontrivial if and only if $\beta(v) < 0$ (Proposition 4.6). The above MMP result implies that the minimum can be approached by a sequence of special divisorial valuations. As a consequence, one can adapt the method developed in [14] to show that there is minimizing valuation which is quasimonomial; see Theorem 4.10. On the other hand, the $\boldsymbol{H}^{\mathrm{NA}}$ invariant for special test configurations is expressed as the $\widetilde{\beta}$ invariant; see Lemma 4.2. Combining these discussions, we will prove (see Sections 2.2 and 2.5 for relevant notation):

Theorem 1.5 *For any $\mathbb{Q}$–Fano variety X, we have the identity*

$$(2) \qquad \inf_{\mathcal{F} \text{ filtration}} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}) = \inf_{(\mathcal{X}, \mathcal{L}, a\eta) \text{ special}} (\boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta)) = \inf_{v \in \mathrm{Val}(X)} \widetilde{\beta}(v).$$

Moreover, the last infimum is achieved by a quasimonomial valuation.

As in the cases of normalized volume, we conjecture that the minimizer is unique and induces a special $\mathbb{R}$–test configuration (see Conjecture 4.11)[2] whose central fiber (with the induced vector field) must then be K–semistable by the following result. When X is smooth, by the result of Dervan and Székelyhidi [31] the existence of such special minimizing valuation is implied by the work of Chen and Wang [29] and Chen, Sun and Wang [28]. We also note that optimal degenerations (of various kinds) in the toric case are well studied; see Wang and Zhu [75] for the toric result for Kähler–Ricci flow.

Theorem 1.6 (Theorem 5.2) *A special $\mathbb{R}$–test configuration minimizes $\boldsymbol{H}^{\mathrm{NA}}$ if and only if its central fiber is K–semistable.*

The uniqueness in Theorem 1.2 about the semistable degeneration is nothing but the result on the uniqueness of the minimizer of $\widetilde{\beta}$ among all quasimonomial valuations associated to special $\mathbb{R}$–test configurations. The proof of this fact uses the technique of initial term degeneration, again motivated by study of normalized volumes; see Li [50] and Li and Xu [58; 57]. This process essentially reduces the question to the uniqueness of minimizer of $\boldsymbol{H}^{\mathrm{NA}}$ (actually a variant of $\boldsymbol{H}^{\mathrm{NA}}$ after the work of Xu and Zhuang [79]) along an interpolation between a fixed filtration and a weight filtration (induced by a holomorphic vector field) on the central fiber. The interpolation is constructed by using the rescaling of

---

[2]This has recently been confirmed by Blum, Liu, Xu and Zhuang [15].

twist of the fixed filtration (the twist of filtration is in the sense of Li [52] generalizing Hisamoto [41]), which we can deal with using the technique of Newton–Okounkov bodies and Boucksom and Jonsson's work on the characterization of asymptotically equivalent filtrations. Our valuative formulation is useful because filtrations associated to valuations are asymptotically equivalent if and only if they are the same (Corollary 2.28 and Lemma 2.29). Again unlike the case of normalized volumes or the case of the δ invariant in Blum, Liu and Zhou [16], the minimizing valuation in the current global setting is expected to be absolutely unique, not just up to rescaling or twisting. This is because of a strict convex property of the $\boldsymbol{H}^{\mathrm{NA}}$–functional, which goes back to Tian and Zhu's work in [71] on the uniqueness of Kähler–Ricci vector fields from the Lie algebra of a torus.

To deal with the polystable step, we first introduce the equivariant version of normalized volumes. Most results about normalized volumes can be generalized for the equivariant version. Finally we complete the proof of Theorem 1.3 by adapting the argument in Li, Wang and Xu [55] about uniqueness of K–polystable degeneration of K–semistable Fano varieties.

To end this introduction, the following table summarizes the quantities used in each of the two steps:

degenerations	semistable	polystable
valuations	$\mathrm{Val}(X)$	$\mathrm{Val}_{C,o}^{\mathbb{C}^* \times \mathbb{T}}$
antiderivative	$\boldsymbol{H}^{\mathrm{NA}}, \widetilde{\beta}$	$\widehat{\mathrm{vol}}_g$
derivative	$\boldsymbol{D}_\xi^{\mathrm{NA}}, \mathrm{Fut}_\xi$	$\boldsymbol{D}_\xi^{\mathrm{NA}}, \beta_g$
derivative formula	(172)	(191)

Postscript   After we finished the paper, we were informed by F Wang and X Zhu that they use analytic methods to prove related uniqueness results for Kähler–Ricci flow limits based on their recent work on the Hamilton–Tian conjecture; see [73; 74].

After this paper appeared, there have been several important developments. In Li and Li [59], based on the minimization setup and uniqueness results in this paper, the authors calculated nontrivial examples of limits of Kähler–Ricci flows on some unstable Fano varieties which are compactifications of homogeneous varieties under some complex reductive group. Very recently, the paper of Blum, Liu, Xu and Zhuang [15] continues and completes the algebraic study of minimization problem proposed in this paper, based on the recent breakthrough on the high-rank finite generation conjecture. In particular, Conjecture 4.11 in our paper is now confirmed in [15].

2 Preliminaries

2.1 Some notation

Let X be a $\mathbb{Q}$–Fano variety. In this paper for the simplicity of notation, we assume that $-K_X$ is Cartier. The modification to the general $\mathbb{Q}$–Cartier case is straightforward; see eg [52]. For any $m, \ell \in \mathbb{N}$, set

$$\text{(3)} \qquad R_m := H^0(X, -mK_X), \qquad R = \bigoplus_{m=0}^{+\infty} R_m,$$

$$\text{(4)} \qquad N_m = \dim R_m, \qquad V = (-K_X)^{\cdot n} = \lim_{m \to +\infty} \frac{N_m}{m^n/n!},$$

$$\text{(5)} \qquad R_m^{(\ell)} := H^0(X, -m\ell K_X), \qquad R^{(\ell)} = \bigoplus_{m=0}^{+\infty} R_m^{(\ell)}.$$

We will denote by $\mathrm{Val}(X)$ the space of real valuations on $\mathbb{C}(X)$, by $\overset{\circ}{\mathrm{Val}}(X)$ the set of real valuations v with $A_X(v) < +\infty$, and by $X_{\mathbb{Q}}^{\mathrm{div}}$ the set of divisorial valuations, ie the valuations of the form $a \cdot \mathrm{ord}_E$ with $a \geq 0$ and E a prime divisor over X. A valuation $v \in \mathrm{Val}(X)$ is quasimonomial if there exist a birational morphism $Y \to X$ and a simple normal crossing divisors $E = \bigcup_{i=1}^{d} E_i \subset Y$ such that v is a monomial valuation on Y with respect to the local coordinates defining E_i, whose center of v over Y is an irreducible component of $\bigcap_{i \in J} E_i$, where $J \subseteq \{1, \ldots, d\}$ is a subset. We denote by $\mathrm{QM}(Y, E)$ the set of such quasimonomial valuations. We refer to [35; 44] for more details about such quasimonomial (or equivalently the Abhyankar) valuations.

In this paper, $\mathbb{T}$ denotes a complex torus $(\mathbb{C}^*)^r = ((S^1)^r)_{\mathbb{C}}$ that acts effectively on a $\mathbb{Q}$–Fano variety X. There is a canonical action of $\mathbb{T}$ on (any multiple of) $-K_X$. Set

$$\text{(6)} \qquad N_{\mathbb{Z}} = \mathrm{Hom}(\mathbb{C}^*, \mathbb{T}), \quad N_{\mathbb{R}} = N_{\mathbb{Z}} \otimes_{\mathbb{Z}} \mathbb{R}, \quad M_{\mathbb{Z}} = \mathrm{Hom}(\mathbb{T}, \mathbb{C}^*), \quad M_{\mathbb{R}} = M_{\mathbb{Z}} \otimes_{\mathbb{Z}} \mathbb{R}.$$

For any $\xi \in N_{\mathbb{R}}$, we have a valuation $\mathrm{wt}_\xi \in \mathrm{Val}(X)$ as follows. For any $f \in \mathbb{C}(X) = \bigoplus_{\alpha \in M_{\mathbb{Z}}} \mathbb{C}(X)_\alpha$,

$$\text{(7)} \qquad \mathrm{wt}_\xi(f) = \min\left\{ \langle \alpha, \xi \rangle \,\Big|\, f = \sum_\alpha f_\alpha, \, f_\alpha \neq 0 \right\}.$$

Moreover, for any $m \in \mathbb{N}$, we have a weight decomposition induced by the canonical $\mathbb{T}$–action on $(X, -mK_X)$:

$$\text{(8)} \qquad R_m = \bigoplus_{\alpha \in M_{\mathbb{Z}}} (R_m)_\alpha = (R_m)_{\alpha_1^{(m)}} \oplus \cdots \oplus (R_m)_{\alpha_{N_m}^{(m)}}.$$

Moreover, we will use the following notation for any $\mathbb{Q}$–Fano variety. Let $e^{-\widetilde{\varphi}}$ be an $(S^1)^r$–invariant smooth positively curved Hermitian metric on $-K_X$ (eg as the restriction of a Fubini–Study metric under

an equivariant embedding of X into projective space). We identify any $\eta \in N_{\mathbb{R}}$ with the corresponding holomorphic vector field on X. Because $\mathbb{T}$–action canonically lifts to an action on $-K_X$, we can set

$$(9) \qquad \theta_{\widetilde{\varphi}}(\eta) = -\frac{\mathfrak{L}_{\eta} e^{-\widetilde{\varphi}}}{e^{-\widetilde{\varphi}}}.$$

Then $\theta_{\widetilde{\varphi}}(\eta)$ is a Hamiltonian function of η with respect to $\mathrm{dd}^c \widetilde{\varphi} = (\sqrt{-1}/2\pi)\partial\bar{\partial}\widetilde{\varphi} \geq 0$:

$$(10) \qquad \iota_{\eta}\mathrm{dd}^c\widetilde{\varphi} = \frac{\sqrt{-1}}{2\pi}\bar{\partial}\theta_{\widetilde{\varphi}}(\eta).$$

Moreover, $(z,\eta) \mapsto \theta_{\widetilde{\varphi}}(\eta)(z)$ is equivalent to the moment map $\boldsymbol{m}_{\widetilde{\varphi}} \colon X \to M_{\mathbb{R}}$ whose image is the moment polytope P of $\mathbb{T}$–action on $(X, -K_X)$ which does not depend on the choice of $\widetilde{\varphi}$. It is known that the measure

$$(11) \qquad \frac{n!}{m^n}\sum_i \dim(R_m)_{\alpha_i^{(m)}} \cdot \delta_{\alpha_i^{(m)}/m}$$

converges weakly to the Duistermaat–Heckman measure $(\boldsymbol{m}_{\widetilde{\varphi}})_*(\mathrm{dd}^c\widetilde{\varphi})^n$; see [23] or [8, Proposition 4.1].

For any subset $S \subseteq \mathbb{R}^n$, we will use dy_S or just dy to denote the Lebesgue measure of S.

2.2 $\mathbb{R}$–test configuration and filtrations

We will use extensively the language of filtrations:

Definition 2.1 [17] A filtration $\mathcal{F} := \mathcal{F}R_{\bullet}$ of the graded $\mathbb{C}$–algebra $R = \bigoplus_{m=0}^{+\infty} R_m$ consists of a family of subspaces $\{\mathcal{F}^{\lambda}R_m\}_x$ of R_m for each $m \geq 0$ with the following properties:

- **Decreasing** $\mathcal{F}^{\lambda}R_m \subseteq \mathcal{F}^{\lambda'}R_m$ if $\lambda \geq \lambda'$.
- **Left continuous** $\mathcal{F}^{\lambda}R_m = \bigcap_{\lambda'<\lambda}\mathcal{F}^{\lambda'}R_m$.
- **Multiplicative** $\mathcal{F}^{\lambda}R_m \cdot \mathcal{F}^{\lambda'}R_{m'} \subseteq \mathcal{F}^{\lambda+\lambda'}R_{m+m'}$ for any $\lambda,\lambda' \in \mathbb{R}$ and $m,m' \in \mathbb{Z}_{\geq 0}$.
- **Linearly bounded** There exist $e_-, e_+ \in \mathbb{Z}$ such that $\mathcal{F}^{me_-}R_m = R_m$ and $\mathcal{F}^{me_+}R_m = 0$ for all $m \in \mathbb{Z}_{\geq 0}$.

Similarly one defines filtration on $R^{(\ell)}$ for any $\ell \geq 1 \in \mathbb{N}$.

Example 2.2 Given any valuation $v \in \overset{\circ}{\mathrm{Val}}(X)$, we have an associated filtration $\mathcal{F} = \mathcal{F}_v$:

$$(12) \qquad \mathcal{F}_v^{\lambda}R_m := \{s \in R_m \mid v(s) \geq \lambda\}.$$

In particular, if there is a $\mathbb{T}$–action on X, for any $\xi \in N_{\mathbb{R}}$, we have a filtration $\mathcal{F}_{\mathrm{wt}_{\xi}}$ associated to the valuation wt_{ξ} in (7).

The trivial filtration $\mathcal{F}_{\mathrm{triv}}$ is the filtration associated to the trivial valuation: $\mathcal{F}_{\mathrm{triv}}^x R_m$ is equal to R_m if $x \leq 0$, and is equal to 0 if $x > 0$.

Example 2.3 For any filtration $\mathcal{F}$, we will denote by $\mathcal{F}_{\mathbb{Z}}$ the filtration defined by $\mathcal{F}_{\mathbb{Z}}^{\lambda} R_m = \mathcal{F}^{\lceil \lambda \rceil} R_m$.

Definition 2.4 [45; 46; 49] We say a valuation $\mathfrak{v} \colon \mathbb{C}(X) \to \mathbb{Z}^n$ (where $\mathbb{Z}^n$ is ordered lexicographically) is a faithful valuation if $\mathfrak{v}(\mathbb{C}(X)) \cong \mathbb{Z}^n$. Note that such a valuation always has at most one-dimensional leaves (in the sense of [45]): if $\mathfrak{v}(f) = \mathfrak{v}(g)$ for $f, g \in \mathbb{C}(X)$, then there exists $c \in \mathbb{C}^*$ satisfying $\mathfrak{v}(f + cg) > \mathfrak{v}(f)$.

Fix such a faithful valuation $\mathfrak{v}$. For any $t \in \mathbb{R}$, define the Newton–Okounkov body of the graded linear series

$$(13) \qquad \mathcal{F}^{(t)} := \mathcal{F}^{(t)} R_\bullet := \{\mathcal{F}^{tm} R_m\}$$

as the closed convex hull of unions of rescaled values of elements from $\mathcal{F}^{(t)}$:

$$(14) \qquad \Delta(\mathcal{F}^{(t)}) = \overline{\bigcup_{m=1}^{+\infty} \frac{1}{m} \mathfrak{v}(\mathcal{F}^{tm} R_m)}.$$

By the theory of Newton–Okounkov bodies [62; 49; 46], we know that

$$(15) \qquad n! \cdot \mathrm{vol}(\Delta(\mathcal{F}^{(t)})) = \mathrm{vol}(\mathcal{F}^{(t)} R_\bullet) = \lim_{m \to +\infty} \frac{\dim_{\mathbb{C}} \mathcal{F}^{mt} R_m}{m^n / n!}.$$

When $t \ll 0$,

$$(16) \qquad \Delta(\mathcal{F}^{(t)}) =: \Delta_{\mathfrak{v}}(X, -K_X) = \Delta(X)$$

is associated to the complete graded linear series $\{R_m\}_m$. Following [17], define the concave transform

$$(17) \qquad G^{\mathcal{F}} \colon \Delta(X) \to \mathbb{R}, \quad G^{\mathcal{F}}(y) = \sup\{t \mid y \in \Delta(\mathcal{F}^{(t)})\}.$$

Given any filtration $\mathcal{F} = \{\mathcal{F}^{\lambda} R_m\}_{\lambda \in \mathbb{R}}$ and $m \in \mathbb{Z}_{\geq 0}$, the successive minima on R_m is the decreasing sequence

$$\lambda_{\max}^{(m)} = \lambda_1^{(m)} \geq \cdots \geq \lambda_{N_m}^{(m)} = \lambda_{\min}^{(m)}$$

defined by

$$\lambda_j^{(m)} = \max\{\lambda \in \mathbb{R} \mid \dim_{\mathbb{C}} \mathcal{F}^{\lambda} R_m \geq j\}.$$

Theorem 2.5 [17] (i) *The function $x \mapsto \mathrm{vol}(\mathcal{F}^{(x)} R_\bullet)^{1/n}$ is concave on $(-\infty, \lambda_{\max})$ and vanishes on $(\lambda_{\max}, +\infty)$.*

(ii) *As $m \to +\infty$, the Dirac-type measure*

$$(18) \qquad \nu_m = \frac{n!}{m^n} \sum_i \delta_{\lambda_i^{(m)}/m} = -\frac{d}{dt} \frac{\dim_{\mathbb{C}} \mathcal{F}^{mt} H^0(Z, m\ell_0 L)}{m^n / n!}$$

converges weakly to a measure with total mass $V = (-K_X)^{\cdot n}$:

$$(19) \qquad \mathrm{DH}(\mathcal{F}) := n! \cdot (G^{\mathcal{F}})_* \, dy = -d\, \mathrm{vol}(\mathcal{F}^{(t)}),$$

where dy is the Lebesgue measure on $\Delta(X)$.

(iii) *The support of the measure* $\mathrm{DH}(\mathcal{F})$ *is given by* $\mathrm{supp}(\mathrm{DH}(\mathcal{F})) = [\lambda_{\min}, \lambda_{\max}]$, *with*

$$\lambda_{\min} := \lambda_{\min}(\mathcal{F}) := \inf\{t \in \mathbb{R} \mid \mathrm{vol}(\mathcal{F}^{(t)}) < V\}, \tag{20}$$

$$\lambda_{\max} := \lambda_{\max}(\mathcal{F}) := \lim_{m \to +\infty} \frac{\lambda_{\max}^{(m)}}{m} = \sup_{m \geq 1} \frac{\lambda_{\max}^{(m)}}{m}. \tag{21}$$

Moreover, $\mathrm{DH}(\mathcal{F})$ *is absolutely continuous with respect to the Lebesgue measure, except perhaps for a point mass at* $\lambda_{\max}$.

Example 2.6 If $v \in \mathrm{Val}(X)$ is quasimonomial, it is shown in [22] that DH is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, ie there is no Dirac mass at $\lambda_{\max}(\mathcal{F}_v)$.

Definition 2.7 Let $\mathcal{F}$ be any filtration. For any $a > 0$ the a–rescaling of $\mathcal{F}$ is given by

$$(a\mathcal{F})^\lambda R_m = \mathcal{F}^{\lambda/a} R_m. \tag{22}$$

For any $b \in \mathbb{R}$, the b–shift is given by

$$\mathcal{F}(b)^\lambda R_m = \mathcal{F}^{\lambda - bm} R_m. \tag{23}$$

Set

$$a\mathcal{F}(b) = (a\mathcal{F})(b) = a(\mathcal{F}(b/a)), \quad \text{ie} \quad a\mathcal{F}(b)^x R_m = \mathcal{F}^{(x-bm)/a} R_m. \tag{24}$$

We have the easy identities

$$\Delta(a\mathcal{F}(b)^{(t)}) = \Delta(\mathcal{F}^{((t-b)/a)}), \quad G_{a\mathcal{F}(b)} = aG_{\mathcal{F}} + b, \quad \mathrm{vol}(a\mathcal{F}(b)^{(t)}) = \mathrm{vol}(\mathcal{F}^{((t-b)/a)}). \tag{25}$$

For any $f_m \in R_m$, set

$$\bar{v}_{\mathcal{F}}(f_m) = \sup\{\lambda \mid f_m \in \mathcal{F}^\lambda R_m\} = \max\{\lambda; f_m \in \mathcal{F}^\lambda R_m\}, \tag{26}$$

and for any $f = \sum_m f_m \in R = \bigoplus_m R_m$ with $f_m \in R_m$, set

$$\bar{v}_{\mathcal{F}}\left(\sum_m f_m\right) = \min\{\bar{v}_{\mathcal{F}}(f_m) \mid f_m \neq 0 \in R_m\}. \tag{27}$$

Then $\bar{v}_{\mathcal{F}}$ is a semivaluation on $R = \bigoplus_m R_m$, satisfying

$$\bar{v}_{\mathcal{F}}(f + g) \geq \min\{\bar{v}_{\mathcal{F}}(f), \bar{v}_{\mathcal{F}}(g)\} \quad \text{and} \quad \bar{v}_{\mathcal{F}}(fg) \geq \bar{v}_{\mathcal{F}}(f) \cdot \bar{v}_{\mathcal{F}}(g). \tag{28}$$

Set

$$\Gamma^+(\mathcal{F}) := \{\lambda_i^{(m)} \mid m \geq 0, 1 \leq i \leq N_m\}. \tag{29}$$

Denote by $\Gamma(\mathcal{F})$ the group of $\mathbb{R}$ generated by $\Gamma^+(\mathcal{F})$.

Definition 2.8   • The extended Rees algebra and associated graded algebra of a filtration $\mathcal{F}$ are
defined as

$$(30) \qquad \mathcal{R}(\mathcal{F}) = \bigoplus_{m \geq 0} \bigoplus_{\lambda \in \Gamma(\mathcal{F})} t^{-\lambda} \mathcal{F}^{\lambda} R_m,$$

$$(31) \qquad \mathrm{Gr}(\mathcal{F}) = \bigoplus_{m \geq 0} \bigoplus_{\lambda \in \Gamma(\mathcal{F})} t^{-\lambda} \mathcal{F}^{\lambda} R_m / \mathcal{F}^{>\lambda} R_m,$$

where $\mathcal{F}^{>\lambda} R_m = \{ f \in R_m \mid v_{\mathcal{F}}(f) > \lambda \}$.

• If $\mathcal{R}(\mathcal{F})$ is finitely generated, we say that $\mathcal{F}$ is finitely generated and call $\mathcal{F}$ an $\mathbb{R}$–test configuration. In this case, $\Gamma(\mathcal{F})$ is a finitely generated free Abelian group: $\Gamma(\mathcal{F}) \cong \mathbb{Z}^{\mathrm{rk}(\mathcal{F})}$ for some positive integer $\mathrm{rk}(\mathcal{F}) \in \mathbb{Z}_{>0}$, and we will call $\mathrm{rk}(\mathcal{F})$ the rank of $\mathcal{F}$. Moreover, $\mathrm{Gr}(\mathcal{F})$ is also finitely generated, and we call the projective scheme $\mathrm{Proj}(\mathrm{Gr}(\mathcal{F})) =: X_{\mathcal{F},0}$ the central fiber of $\mathcal{F}$.

There is an induced filtration $\mathcal{F}|_{X_{\mathcal{F},0}} := \mathcal{F}' R' = \{\mathcal{F}' R_m'\}$ on $R' := \mathrm{Gr}(\mathcal{F})$, the homogeneous coordinate ring of the central fiber:

$$(32) \qquad \mathcal{F}'^{\lambda} R_m' = \bigoplus_{\lambda_i^{(m)} \geq \lambda} \mathcal{F}^{\lambda_i^{(m)}} R_m / \mathcal{F}^{>\lambda_i^{(m)}} R_m.$$

The $\Gamma(\mathcal{F})$ grading of $\mathrm{Gr}(\mathcal{F})$ corresponds to a holomorphic vector field $\eta = \eta_{\mathcal{F}}$ on the central fiber, which generates an action by a complex torus of dimension $\mathrm{rk}(\mathcal{F})$.

• We say an $\mathbb{R}$–test configuration $\mathcal{F}$ is special if its central fiber $X_{\mathcal{F},0}$ is a $\mathbb{Q}$–Fano variety and there is an isomorphism $\mathrm{Gr}(\mathcal{F}) \cong R(X_{\mathcal{F},0}, -K_{X_{\mathcal{F},0}}) =: R'$. In this case, there is a $\sigma \in \mathbb{R}$ such that

$$(33) \qquad \mathcal{F}' R' = \mathcal{F}'_{\mathrm{wt}_\eta} R'(-\sigma).$$

Remark 2.9   We can naturally extend the above definition to filtrations on $R^{(\ell)}$ for any $\ell \in \mathbb{N}_{\geq 1}$. Indeed we will actually identify two filtrations if they induce the same non-Archimedean metric on $(X^{\mathrm{NA}}, L^{\mathrm{NA}})$ with $L = -K_X$. See Definition 2.17.

There are two equivalent geometric descriptions of $\mathbb{R}$–test configurations, which we now explain.

(I) Geometric $\mathbb{R}$–TC I   Let $\iota : X \to \mathbb{P}^{N_\ell}$ be a Kodaira embedding by a basis of $R_\ell = H^0(X, \ell(-K_X))$ for some $\ell > 0$, and let η be a holomorphic vector field on $\mathbb{P}^{N_\ell - 1} = \mathbb{P}(H^0(X, \ell(-K_X)^*))$ that generates an effective holomorphic action on $\mathbb{P}^{N_\ell - 1}$ by a torus $\mathbb{T}$ of rank r. Then we get a weight decomposition $R_\ell = \bigoplus_{\alpha \in \mathbb{Z}^r} R_{\ell,\alpha}$ and a filtration on R_ℓ by setting

$$(34) \qquad \mathcal{F}^{\lambda} R_\ell = \bigoplus_{\langle \alpha, \eta \rangle \geq \lambda} R_{\ell,\alpha}.$$

The filtration $\mathcal{F} R_\ell$ generates a filtration on $\mathcal{F} R^{(\ell)}$, which is an $\mathbb{R}$–test configuration $\mathcal{F}$. The following lemma generalizes the well-known fact for test configurations; see [33; 76; 64].

Lemma 2.10 *Any* $\mathbb{R}$*–test configuration, which by definition is a finitely generated filtration, is obtained in this way.*

Proof To see this, we assume again that $\mathcal{F}$ is generated by $\mathcal{F}R_\ell$. For simplicity of notation, set $V = R_\ell$ and $\lambda_i = \lambda_i^{(\ell)}$. By shifting the filtration, we can normalize $\lambda_{N_\ell} = 0$ and assume that we have the relation

$$\lambda_1 = \cdots = \lambda_{i_1} =: w_1 > \lambda_{i_1+1} = \cdots = \lambda_{i_2} =: w_2$$
$$\vdots$$
$$> \lambda_{i_{k-2}+1} = \cdots = \lambda_{i_{k-1}} =: w_{k-1}$$
$$> \lambda_{i_{k-1}+1} = \cdots = \lambda_{N_\ell} =: w_k = 0.$$

In other words, $\{w_1, \ldots, w_k\}$ is the set of distinct values of successive minima and we have a usual filtration,

$$\text{(35)} \qquad \{0\} \subsetneq \mathcal{F}^{w_1} V \subsetneq \mathcal{F}^{w_2} V \subsetneq \cdots \subsetneq \mathcal{F}^{w_k} = V.$$

In other words, we can equivalently describe an $\mathbb{R}$–filtration by the language of weighted flags. Fixing a reference Hermitian inner product H_0 on $V = R_\ell$, we can assign to the flag (35) a decomposition

$$\text{(36)} \qquad V = V_1 \oplus V_2 \oplus \cdots \oplus V_k,$$

where $V_1 = \mathcal{F}^{w_1} V$ and V_j is the H_0–orthogonal complement of $\mathcal{F}^{w_{j-1}} V$ inside $\mathcal{F}^{w_j} V$, which has dimension $i_j - i_{j-1} =: d_j$.

Fix a maximal $\mathbb{Q}$–linearly independent subset of $\{w_1, \ldots, w_k\}$ to be

$$\text{(37)} \qquad 0 > w_2 =: \zeta_1 > \cdots > w_{p_r} =: \zeta_r.$$

So for each w_j we can find a vector of rational numbers $\vec{r}_j = (r_{j1}, \ldots, r_{jr}) \in \mathbb{Q}$ such that $w_j = \sum_{p=1}^r r_{jp} \zeta_p$. Finding a common multiple D of the denominators of $\{r_{jp} \mid 1 \le j \le k, 1 \le p \le r\}$, we set $\eta = \zeta/D$ and $\alpha_j = D\vec{r}_j$, so that

$$\text{(38)} \qquad w_j = \sum_{p=1}^r \alpha_{jp} \zeta_p = \langle \alpha_j, \eta \rangle.$$

In this way we get a $(\mathbb{C}^*)^r$ representation V, whose weight decomposition is given by (36), where V_j consists of elements of weight α_j, and

$$\mathcal{F}^\lambda V = \bigoplus_{\langle \alpha_j, \eta \rangle \ge a} V_j = \left\{ v = \sum_{j=1}^k v_j \; \Big| \; \min\{\langle \alpha_j, \eta \rangle \mid v_j \ne 0\} \ge \lambda \right\}. \qquad \square$$

From another point of view, let $\mathcal{I}_X \subset \mathbb{C}[Z_1, \ldots, Z_{N_\ell}] = S$ be the homogeneous ideal of X. For each $d \in \mathbb{N}$, the $\mathbb{T}$–action induces a representation of $\mathbb{T}$ on S_d, the set of degree-d homogeneous polynomials. The holomorphic vector field η induces an order on the weights of these $\mathbb{T}$–representations. Choosing a set of homogeneous generators of $\mathcal{I}_X$, the initial term with respect to this order generates the ideal

of $X_{\mathcal{F},0}$. If σ_η denotes the one-parameter $\mathbb{R}$–group generated by η, we have the convergence of algebraic cycles (or schemes)

$$\lim_{s \to +\infty} \sigma_\eta(s) \circ [X] = [X_{\mathcal{F},0}].\tag{39}$$

So we say that the $\mathbb{R}$–action generated by η degenerates X into a projective scheme $X_{\mathcal{F},0}$.

By perturbing $\eta \in N_\mathbb{R}$, we can find a sequence of rational vector $\eta_k \in N_\mathbb{Q}$ converging to η. For $k \gg 1$, η_k induces an $\mathbb{R}$–test configuration of rank one with the same central fiber $X_{\mathcal{F},0}$.

(II) **Geometric $\mathbb{R}$–TC II** This description is essentially contained in [66, Section 2]. For any $\mathbb{R}$–test configuration, we set $B = \mathrm{Spec}(\mathbb{C}(\Gamma^+(\mathcal{F}))) \cong \mathbb{C}^r$. Then there is a flat family

$$\mathcal{X} = \mathrm{Proj}_{\mathbb{C}^r}(\mathcal{R}(\mathcal{F})) \to B\tag{40}$$

such that the generic fiber is isomorphic to X and a special fiber isomorphic to $X_{\mathcal{F},0}$. Set $\mathcal{L}$ to be the relative ample line bundle $\mathcal{O}_{\mathcal{X}/\mathbb{C}^r}(1)$. Fix $m \geq 0$. For any $\lambda \in \mathbb{R}$, we set $\lceil \lambda \rceil = \min\{\lambda_i^{(m)} \mid \lambda_i^{(m)} \geq \lambda\} = \langle \alpha, \eta_{\mathcal{F}} \rangle$ for $\alpha \in M_\mathbb{Z}$. Then for any $\tau = (\tau_1, \ldots, \tau_r) \in \mathbb{C}^r$, we set $\tau^{-\lceil \lambda \rceil} = \prod_{i=1}^r \tau_i^{\alpha_i}$, to get

$$\mathcal{F}^\lambda R_m = \{s \in R_m \mid \tau^{-\lceil \lambda \rceil}\overline{s} \text{ extends to a holomorphic section of } m\mathcal{L} \to \mathcal{X}\},\tag{41}$$

where $\overline{s}$ is the meromorphic section of $m\mathcal{L}$ defined as the pullback of s via the projection $(\mathcal{X}, \mathcal{L}) \times_B (\mathbb{C}^*)^r \cong (X, -K_X) \times (\mathbb{C}^*)^r \to X$.

Lemma 2.11 *If $\mathrm{Gr}(\mathcal{F})$ is an integral domain, then the semivaluation $\overline{v}_{\mathcal{F}}$ in (26) defines a valuation on the quotient field of R. Denote by $v_{\mathcal{F}}$ the restriction of $\overline{v}_{\mathcal{F}}$ to $\mathbb{C}(X)$: for $f = s_1/s_2 \in \mathbb{C}(X)$ with $s_1, s_2 \in R_m$, set*

$$v_{\mathcal{F}}(f) = \overline{v}_{\mathcal{F}}(s_1) - \overline{v}_{\mathcal{F}}(s_2).\tag{42}$$

Then there exists $\sigma > 0$ such that $\mathcal{F} = \mathcal{F}_{v_{\mathcal{F}}}(-\sigma)$. In particular, this statement applies to any special $\mathbb{R}$–test configuration.

Proof Fix any two homogeneous elements $s_i \in R_{m_i}$ for $i = 1, 2$. Assume that $\overline{v}_{\mathcal{F}}(f_i) = s_i$. Then $s_i' \in R'_{m_i, x_i}$. Because $\mathrm{Gr}(\mathcal{F})$ is integral, $s_1' s_2' \neq 0 \in R'_{m_1+m_2, x_1+x_2}$, which implies that $\overline{v}_{\mathcal{F}}(s_1 s_2) = x_1 + x_2 = \overline{v}_{\mathcal{F}}(s_1) + \overline{v}_{\mathcal{F}}(s_2)$. From this, we easily see that $\overline{v}_{\mathcal{F}}$ is a real valuation.

Assume $f = s_1/s_2 = \tilde{s}_1/\tilde{s}_2$. Then $s_1 \cdot \tilde{s}_2 = s_2 \cdot \tilde{s}_1$ and hence $\overline{v}_{\mathcal{F}}(s_1) - \overline{v}_{\mathcal{F}}(s_2) = \overline{v}_{\mathcal{F}}(\tilde{s}_1) - \overline{v}_{\mathcal{F}}(\tilde{s}_2)$. So $v_{\mathcal{F}}$ in (42) is well defined.

For any $s_i \neq 0 \in R_m$ with $i = 1, 2$, by construction $\overline{v}_{\mathcal{F}}(s_1) - v_{\mathcal{F}}(s_1) = \overline{v}_{\mathcal{F}}(s_2) - \overline{v}(s_2)$. This means $b_m := v_{\mathcal{F}} - \overline{v}_{\mathcal{F}}$ is constant on $R_m \setminus \{0\}$. It is easy to see that $\sigma_{m_1}\sigma_{m_2} = \sigma_{m_1+m_2}$. So we can set $\sigma = \sigma_m/m$ to get the conclusion. $\square$

An $\mathbb{R}$–test configuration with $\mathrm{rk}(\mathcal{F}) = 1$ is, up to rescaling, associated to the usual test configuration, a notion that plays a basic role in the subject of K–stability.

Definition 2.12 [67; 32; 56] A test configuration of (X, L) is a triple $(\mathcal{X}, \mathcal{L}, \eta)$, sometimes just denoted by $(\mathcal{X}, \mathcal{L})$, that consists of

- a variety $\mathcal{X}$ admitting a $\mathbb{C}^*$–action generated by a holomorphic vector field η and a $\mathbb{C}^*$–equivariant morphism $\pi : \mathcal{X} \to \mathbb{C}$, where the action of $\mathbb{C}^*$ on $\mathbb{C}$ is given by the standard multiplication generated by $-t\partial_t$, and

- a $\mathbb{C}^*$–equivariant π–semiample $\mathbb{Q}$–Cartier $\mathbb{Q}$–divisor $\mathcal{L}$ on $\mathcal{X}$ such that there is an $\mathbb{C}^*$–equivariant isomorphism $i_\eta : (\mathcal{X}, \mathcal{L})|_{\pi^{-1}(\mathbb{C}\setminus\{0\})} \cong (X, L) \times \mathbb{C}^*$.

We denote by $(\bar{\mathcal{X}}, \bar{\mathcal{L}})$ the natural compactification of $(\mathcal{X}, \mathcal{L})$ obtained by adding a trivial fiber at infinity using the isomorphism i_η.

$(X_{\mathbb{C}}, (-K_X)_{\mathbb{C}}, \eta_{\text{triv}}) := (X \times \mathbb{C}, -K_X \times \mathbb{C}, -t\partial_t)$ is called the trivial test configuration. $(\mathcal{X}, \mathcal{L}, \eta)$ is a normal test configuration if $\mathcal{X}$ is a normal variety.

A normal test configuration $(\mathcal{X}, \mathcal{L}, \eta)$ is a special test configuration (resp. weakly special) if $(\mathcal{X}, \mathcal{X}_0)$ is plt (resp. if $(\mathcal{X}, \mathcal{X}_0)$ is log canonical) and $\mathcal{L} = -K_{\mathcal{X}} + c\mathcal{X}_0$ for some $c \in \mathbb{Q}$. By inversion of adjunction, $(\mathcal{X}, \mathcal{L}, \eta)$ being special is equivalent to the condition that $(\mathcal{X}_0, -K_{\mathcal{X}_0})$ is $\mathbb{Q}$–Fano.

Two test configurations $(\mathcal{X}_i, \mathcal{L}_i)$ for $i = 1, 2$ are equivalent if there exists a test configuration $(\mathcal{X}', \mathcal{L}')$ and two $\mathbb{C}^*$–equivariant birational morphisms $\rho_i : \mathcal{X}' \to \mathcal{X}_i$ such that $\rho_1^* \mathcal{L}_1 = \mathcal{L}' = \rho_2^* \mathcal{L}_2$.

Assume that $\mathbb{G}$ is a reductive complex Lie group acting on (X, L). A $\mathbb{G}$–equivariant test configuration of (X, L) is a test configuration $(\mathcal{X}, \mathcal{L}, \eta)$ with the following property:

- There is a $\mathbb{G}$–action on $(\mathcal{X}, \mathcal{L})$ that commutes with the $\mathbb{C}^*$–action generated by η and the action of $\mathbb{G}$ on $(\mathcal{X}, \mathcal{L}) \times_{\mathbb{C}} \mathbb{C}^* \overset{i_\eta}{\cong} (X, L) \times \mathbb{C}^*$ coincides with the fiberwise action of $\mathbb{G}$ on (the first factor of) $(X, L) \times \mathbb{C}^*$.

As mentioned above, by the work of [76; 65; 19], for any $\mathbb{R}$–test configuration $\mathcal{F}$ with $\text{rk}(\mathcal{F}) = 1$, there exists a test configuration $(\mathcal{X}, \mathcal{L}, \eta)$ and $a > 0$, such that $\Gamma(\mathcal{F}) \cong a\mathbb{Z}$ and $\mathcal{F} = a\mathcal{F}_{(\mathcal{X},\mathcal{L},\eta)}$. In this case, we will also denote the $\mathbb{R}$–test configuration $\mathcal{F}$ by $(\mathcal{X}, \mathcal{L}, a\eta)$ and set

(43)
$$\mathcal{F}_{(\mathcal{X},\mathcal{L},a\eta)} := a\mathcal{F}_{(\mathcal{X},\mathcal{L},\eta)}.$$

The identity (40) becomes

(44)
$$\mathcal{X} = \text{Proj}_{\mathbb{C}[t]}\left(\bigoplus_{m\geq 0}\bigoplus_{j\in\mathbb{Z}} t^{-aj}\mathcal{F}^j R_m\right).$$

Conversely, assume $(\mathcal{X}, \mathcal{L})$ is a test configuration of $(X, L := -K_X)$. Then we associate to it a filtration $\mathcal{F} = \mathcal{F}_{(\mathcal{X},\mathcal{L})}$ as in (41); so $s \in \mathcal{F}^\lambda R_m$ if and only if $t^{-\lceil\lambda\rceil}\bar{s}$ extends to a holomorphic section of $m\mathcal{L}$. In particular, such a construction sets up a one-to-one correspondence between test configurations $(\mathcal{X}, \mathcal{L})$ with ample $\mathcal{L}$, and $\mathbb{R}$–test configurations $\mathcal{F}$ with $\Gamma(\mathcal{F}) \subseteq \mathbb{Z}$; see [19, Proposition 2.15].

Now assume that $(\mathcal{X}, \mathcal{L})$ is normal and there is a $\mathbb{C}^*$–equivariant birational morphism $\rho \colon \mathcal{X} \to X_{\mathbb{C}} := X \times \mathbb{C}$. Write $\mathcal{L} = \rho^* L_{\mathbb{C}} + D$, where $L_{\mathbb{C}} = p_1^* L$. Then by [19, Lemma 5.17], the filtration $\mathcal{F}$ has the more explicit description

$$(45) \qquad \mathcal{F}^\lambda R_m = \bigcap_E \{s \in H^0(X, mL) \mid r(\mathrm{ord}_E)(s) + m\ell_0 \, \mathrm{ord}_E(D) \geq x b_E\},$$

where E runs over the irreducible components of the central fiber $\mathcal{X}_0$, and $b_E = \mathrm{ord}_E(\mathcal{X}_0) = \mathrm{ord}_E(t)$ while $r(\mathrm{ord}_E)$ denotes the restriction of ord_E to $\mathbb{C}(Z)$ under the inclusion $\mathbb{C}(Z) \subset \mathbb{C}(X \times \mathbb{C}^*) = \mathbb{C}(\mathcal{X})$.

When $\mathcal{F} = \mathcal{F}_{(\mathcal{X}, -K_{\mathcal{X}}, \eta)}$ is associated to a special test configuration, Lemma 2.11 applies. In fact, by [19], $v_{\mathcal{F}} = v_{\mathcal{X}_0} = r(\mathrm{ord}_{\mathcal{X}_0})$ and by [50], $\sigma = A_X(v_{\mathcal{X}_0})$, so $\mathcal{F}_{(\mathcal{X}, -K_{\mathcal{X}}, \eta)} = \mathcal{F}_{v_{\mathcal{X}_0}}(-A(v_{\mathcal{X}_0}))$. As a consequence, for any $a > 0$, by (24) we have the identity

$$(46) \qquad \mathcal{F}_{(\mathcal{X}, -K_{\mathcal{X}}, a\eta)} = \mathcal{F}_{a v_{\mathcal{X}_0}}(-A(a v_{\mathcal{X}_0})).$$

Note that following Definition 2.8, for any $a > 0$ we say that $(\mathcal{X}, \mathcal{L}, a\eta)$ is a special (resp. normal) $\mathbb{R}$–test configuration if $(\mathcal{X}, \mathcal{L}, \eta)$ is a special (resp. normal) test configuration.

Note that we use the negative sign $-t \partial_t$ in our Definition 2.12. This sign convention will be convenient for our subsequent computations, as illustrated in the following simple example.

Example 2.13 Consider the product test configuration $(\mathcal{X}, \mathcal{L})$ of $(\mathbb{P}^1, \mathcal{O}_{\mathbb{P}^1}(1))$ induced by the $\mathbb{C}^*$–action

$$t \circ [Z_0, Z_1] = [Z_0, t Z_1].$$

Let s_i for $i = 0, 1$ be two holomorphic sections of $H^0(\mathbb{P}^1, \mathcal{O}(1))$ corresponding to the homogeneous coordinates Z_i for $i = 0, 1$. Then t acts on the holomorphic sections by $t \cdot s_0 = s_0$ and $t \circ s_1 = t^{-1} s_1$. The corresponding filtration is given by

$$(47) \qquad \mathcal{F}^\lambda R_m = \mathrm{Span}\{s_0^{m-i} s_1^i \mid 0 \geq -i \geq \lambda\};$$

cf (34). The natural compactification $\bar{\mathcal{X}}$ can be identified with the Hirzebruch surface $\mathbb{P}(\mathcal{O}_{\mathbb{P}^1}(1) \oplus \mathcal{O}_{\mathbb{P}^1})$, and $\bar{\mathcal{L}}$ is given by $\mathcal{O}_{\bar{\mathcal{X}}}(D_\infty)$, where D_∞ is the divisor at infinity; see [56, Example 3]. The successive minima are given by $\{\lambda_i^{(m)}\} = \{-m, -m+1, \dots, 0\}$. In particular, we have

$$(48) \qquad \sum_i \lambda_i^{(m)} = -\tfrac{1}{2} m^2 - \tfrac{1}{2} m = \tfrac{1}{2} \bar{\mathcal{L}}^2 m^2 + \left(\tfrac{1}{2} K_{\bar{\mathcal{X}}}^{-1} \cdot \bar{\mathcal{L}} - 1\right) m.$$

Moreover, $\eta = -z \, \partial/\partial z$, whose Hamiltonian function is given by $\theta(\eta) = -|Z_1|^2/(|Z_1|^2 + |Z_2|^2)$. Note that $\theta(\eta)_* \omega_{\mathrm{FS}} = dy_{[-1,0]} = \mathrm{DH}(\mathcal{F})$.

Example 2.14 If $\mathcal{F}$ is an $\mathbb{R}$–test configuration, then $a\mathcal{F}(b)$ is an $\mathbb{R}$–test configurations for any $(a, b) \in \mathbb{R}_{>0} \times \mathbb{R}$.

Assume $\mathcal{F} = \mathcal{F}_{(\mathcal{X}, \mathcal{L}, \eta)}$ for a test configuration $(\mathcal{X}, \mathcal{L}, \eta)$. Then as mentioned above, for simplicity of notation we will identify $a\mathcal{F}(b)$ with the data $(\mathcal{X}, \mathcal{L} + b\mathcal{X}_0, a\eta)$.

For any $d > 0 \in \mathbb{N}$, we can consider the normalization of the base change,

(49) $$(\mathcal{X}, \mathcal{L}, \eta)^{(d)} := ((\mathcal{X}, \mathcal{L}, \eta) \times_{\mathbb{C}, t \to t^d} \mathbb{C})^{\mathrm{norm}} =: (\mathcal{X}^{(d)}, \mathcal{L}^{(d)}, \eta^{(d)}).$$

On the other hand, $\mathbb{Z}_d = \langle e^{2\pi \sqrt{-1}/d} \rangle \hookrightarrow \mathbb{C}^*$ naturally acts on the $(\mathcal{X}, \mathcal{L})$ and we can take a quotient

(50) $$(\mathcal{X}, \mathcal{L}, \eta)/_{\mathbb{Z}_d} = (\mathcal{X}^{(1/d)}, \mathcal{L}^{(1/d)}, \eta^{(1/d)})$$

to get a test configuration with a nonreduced central fiber in general.

With this notation, for any $a > 0 \in \mathbb{Q}$ we then have the natural identification

(51) $$\mathcal{F}_{(\mathcal{X}, \mathcal{L}, \eta)^{(a)}} = a \cdot \mathcal{F}_{(\mathcal{X}, \mathcal{L}, \eta)} = \mathcal{F}_{(\mathcal{X}, \mathcal{L}, a\eta)}.$$

For a filtration $\mathcal{F}R_\bullet$, choose e_- and e_+ as in Definition 2.1. For convenience, we can choose $e_+ = \lceil \lambda_{\max}(\mathcal{F}R) \rceil \in \mathbb{Z}$. Set $e = e_+ - e_-$ and define (fractional) ideals

(52) $\quad I_{m,\lambda} := I_{m,\lambda}^{\mathcal{F}} := \mathrm{Image}(\mathcal{F}^\lambda R_m \otimes \mathcal{O}_X(-mL) \to \mathcal{O}_X),$

(53) $\quad \widetilde{\mathcal{I}}_m := \widetilde{\mathcal{I}}_m^{\mathcal{F}} := I_{(m, me_+)}^{\mathcal{F}} t^{-me_+} + I_{(m, me_+ - 1)}^{\mathcal{F}} t^{1 - me_+} + \cdots + I_{(m, me_- + 1)}^{\mathcal{F}} t^{-me_- - 1} + \mathcal{O}_X \cdot t^{-me_-},$

(54) $\quad \mathcal{I}_m := \mathcal{I}_m^{\mathcal{F}(e_+)} = \widetilde{\mathcal{I}}_m^{\mathcal{F}} \cdot t^{me_+} = I_{(m, me_+)}^{\mathcal{F}} + I_{(m, me_+ - 1)}^{\mathcal{F}} t^1 + \cdots + I_{(m, me_- + 1)}^{\mathcal{F}} t^{me - 1} + (t^{me}) \subseteq \mathcal{O}_{X_{\mathbb{C}}}.$

Definition–Proposition 2.15 [36, Lemma 4.6] *With the above notation, for m sufficiently divisible, define the mth approximating test configuration $(\check{\mathcal{X}}_m^{\mathcal{F}}, \check{\mathcal{L}}_m^{\mathcal{F}})$ as follows:*

(i) *$\check{\mathcal{X}}_m^{\mathcal{F}}$ is the normalization of blowup of $X \times \mathbb{C}$ along the ideal sheaf $\mathcal{I}_m^{\mathcal{F}(e_+)}$.*

(ii) *The semiample $\mathbb{Q}$–divisor is given by*

(55) $$\check{\mathcal{L}}_m^{\mathcal{F}} = \pi^*((-K_X) \times \mathbb{C}) - \frac{1}{m} E_m + e_+ \check{\mathcal{X}}_0,$$

where E_m is the exceptional divisor of the normalized blowup.

For simplicity of notation, we also denote the data by $(\check{\mathcal{X}}_m, \check{\mathcal{L}}_m)$ if the filtration is clear.

It is easy to see that the filtration $\mathcal{F}_{(\check{\mathcal{X}}_m, \check{\mathcal{L}}_m)}$ on $R^{(m)}$ is induced by $\mathcal{F}_{\mathbb{Z}} R_m$ under the canonical map $S^k R_m \to R_{km}$. By [20, Proof of Theorem 4.13], we have the following approximation result.

Proposition 2.16 [20, Proof of Theorem 4.13] *With notation as in Definition–Proposition 2.15, the Duistermaat–Heckmann measures $\mathrm{DH}(\check{\mathcal{X}}_m, \check{\mathcal{L}}_m)$ converge weakly to $\mathrm{DH}(\mathcal{F})$ as $m \to +\infty$.*

Following Boucksom and Jonsson, it is very convenient to use the non-Archimedean metric defined by filtrations. Any filtration (in the sense of Definition 2.1) defines a non-Archimedean metric on $L^{\mathrm{NA}} \to X^{\mathrm{NA}}$. If we denote by ϕ_{triv} the non-Archimedean metric associated to the trivial filtration, then any non-Archimedean metric ϕ on L^{NA} is represented by the real valued function $\phi - \phi_{\mathrm{triv}}$ on $X_{\mathbb{Q}}^{\mathrm{div}}$.

Definition 2.17 Let $\mathcal{F} = \mathcal{F}R_\bullet$ be a filtration. For any $w \in \mathring{\mathrm{Val}}(X)$, define the non-Archimedean metric associated to $\mathcal{F}$ by

$$(56) \qquad (\phi_m^{\mathcal{F}} - \phi_{\mathrm{triv}})(w) = -\frac{1}{m}G(w)(\widetilde{\mathcal{I}}_m^{\mathcal{F}}) = -\frac{1}{m}G(w)(\mathcal{I}_m^{\mathcal{F}(e_+)}t^{-me_+}) = -\frac{1}{m}G(w)(\mathcal{I}_m^{\mathcal{F}(e_+)}) + e_+,$$

$$(57) \qquad (\phi^{\mathcal{F}} - \phi_{\mathrm{triv}})(w) = -G(w)(\widetilde{\mathcal{I}}_\bullet^{\mathcal{F}}) = \lim_{m \to +\infty} \phi_m^{\mathcal{F}}(w).$$

In particular, if $v \in \mathring{\mathrm{Val}}(Z)$ and $\mathcal{F} = \mathcal{F}_v$, then we write $\phi_v = \phi^{\mathcal{F}_v}$.

Note that $\phi_m^{\mathcal{F}} = \phi_{\mathcal{F}(\check{\mathcal{X}}_m, \check{\mathcal{L}}_m)}$ converges to ϕ as $m \to +\infty$. Moreover, if (for simplicity) we assume that $S^k R_m \to R_{km}$ is surjective for all $k, m \geq 1$, then it is an increasing sequence in the sense that if $m_1 \mid m_2$, then $\phi_{m_1}^{\mathcal{F}} \leq \phi_{m_2}^{\mathcal{F}}$. If $\phi^{\mathcal{F}}$ is continuous, then ϕ_m converges to ϕ uniformly by Dini's theorem.

The following transformation rule can be easily verified.

Lemma 2.18 *For any filtration $\mathcal{F}$ and any $(a, b) \in \mathbb{R}_{>0} \times \mathbb{R}$ and $v \in X_\mathbb{Q}^{\mathrm{div}}$,*

$$(58) \qquad (\phi_{a\mathcal{F}(b)} - \phi_{\mathrm{triv}})(v) = a(\phi_{\mathcal{F}} - \phi_{\mathrm{triv}})\left(\frac{v}{a}\right) + b.$$

2.3 Twist of filtrations

Let $\mathcal{F} = \mathcal{F}R_\bullet$ be a $\mathbb{T}$–equivariant filtration, which means that $\mathcal{F}^\lambda R_m$ is a $\mathbb{T}$–invariant subspace of R_m for any $x \in \mathbb{R}$. For $\alpha \in M_\mathbb{Z} = N_\mathbb{Z}^\vee$, denote the weight space by

$$(59) \qquad (R_m)_\alpha = \{s \in R_m \mid \tau \circ s = \tau^\alpha s \text{ for all } \tau \in (\mathbb{C}^*)^r\}.$$

Then we have

$$(60) \qquad (\mathcal{F}^\lambda R_m)_\alpha := \{s \in \mathcal{F}^\lambda R_m \mid \tau \circ s = \tau^\alpha s\} = \mathcal{F}^\lambda R_m \cap (R_m)_\alpha,$$

and the decomposition

$$(61) \qquad \mathcal{F}^\lambda R_m = \bigoplus_{\alpha \in M_\mathbb{Z}} (\mathcal{F}^\lambda R_m)_\alpha.$$

Definition 2.19 [52] For any $\xi \in N_\mathbb{R}$, the ξ–twist of $\mathcal{F}$ is the filtration $\mathcal{F}_\xi R_\bullet$ defined by

$$(62) \qquad \mathcal{F}_\xi^\lambda R_m = \bigoplus_{\alpha \in M_\mathbb{Z}} (\mathcal{F}_\xi^\lambda R_m)_\alpha, \quad \text{where } (\mathcal{F}_\xi^\lambda R_m)_\alpha := (\mathcal{F}^{\lambda - \langle \alpha, \xi \rangle} R_m)_\alpha.$$

Example 2.20 If $\mathcal{F}$ is a $\mathbb{T}$–equivariant $\mathbb{R}$–test configuration, then $\mathcal{F}_\xi$ is also an $\mathbb{R}$–test configuration.

If $\mathcal{F} = \mathcal{F}_{(\mathcal{X}, \mathcal{L}, a\eta)}$ for a test configuration, then we can identify the data $\mathcal{F}_\xi$ with the data $(\mathcal{X}, \mathcal{L}, a\eta + \xi)$; see [41]. If $\xi \in N_\mathbb{Z}$, then $(\mathcal{X}, \mathcal{L}, a\eta + \xi)$ is equivalent to the birational image of the $(\mathcal{X}, \mathcal{L})$ via the birational transform $\sigma_\xi \colon \mathcal{X} \dashrightarrow \mathcal{X}$, $(z, t) \to (\sigma_\xi(t) \cdot z, t)$; see [52].

Moreover, if we start with the trivial filtration $\mathcal{F}_{\mathrm{triv}} = \mathcal{F}_{(X_\mathbb{C}, (-K_X)_\mathbb{C}, -t\partial_t)}$, then $(\mathcal{F}_{\mathrm{triv}})_\xi$ is equal to $\mathcal{F}_{\mathrm{wt}_\xi}$.

Definition 2.21 We say that a faithful valuation $\mathfrak{v}$ in the sense of Definition 2.4 is adapted to the torus action if for any $f \in \mathbb{C}(X)_\alpha$ we have $\mathfrak{v}(f) = (\alpha, \mathfrak{v}^{r+1}(f), \ldots, \mathfrak{v}^n(f)) \in \mathbb{Z}^r \times \mathbb{Z}^{n-r}$.

There always exists a faithful valuation that is adapted to the torus action. This can be constructed as follows. First we choose a $\mathbb{T}$–invariant Zariski-open set U of X as in [3]. Then by the theory of affine T–varieties as developed in [2], there exists a variety Y of dimension $n - r$ and a polyhedral divisor $\mathfrak{D}$ such that

(63)
$$U = \mathrm{Spec}_{\alpha \in M_\mathbb{Z}} H^0(Y, \mathcal{O}(\mathfrak{D}(\alpha))).$$

We can choose a faithful valuation $\mathfrak{v}_Y$ on Y (for example via a flag of varieties as in [49]) and define, for any $f \in H^0(Y, \mathcal{O}(\mathfrak{D}(\alpha))$,

(64)
$$\mathfrak{v}(f) = (\alpha, \mathfrak{v}_Y(f)).$$

Let $\mathfrak{v}$ be such a valuation and $\Delta = \Delta_\mathfrak{v}(X, -K_X) \subset \mathbb{R}^n$ be the associated Newton–Okounkov body. If $p : \mathbb{R}^n = \mathbb{R}^r \times \mathbb{R}^{n-r} \to \mathbb{R}^r$ denotes the natural projection, then we have

(65)
$$p(\Delta) = P = \text{moment map of the } \mathbb{T}\text{–action on } (X, -K_X).$$

The following lemma was already observed in [81], in which a faithful valuation adapted to the torus action was constructed using equivariant infinitesimal flags in the sense of [49]. Here we give a different and direct proof for the reader's convenience.

For simplicity of notation, we write $y = (y_1, \ldots, y_n) = (y', y'') \in \mathbb{R}^r \times \mathbb{R}^{n-r}$ and set

(66)
$$\langle y', \xi \rangle = \sum_{i=1}^r y_i' \xi^i =: \langle y, \xi \rangle.$$

In the last identity, we identify $\xi \in N_\mathbb{R} = \mathbb{R}^r$ with $(\xi, 0) \in \mathbb{R}^n$.

Lemma 2.22 [81] *If $\mathfrak{v}$ is a $\mathbb{Z}^n$–valued valuation adapted to the torus action, then for any $y \in \Delta(-K_X)$,*

(67)
$$G_{\mathcal{F}_\xi}(y) = G_\mathcal{F}(y) + \langle y', \xi \rangle.$$

Proof For any $t > G^\mathcal{F}(y) = \lambda$, there exists $\epsilon > 0$ such that $y \notin \Delta(\mathcal{F}^{(t-\epsilon)})$. Let $\delta_1 = \mathrm{dist}(y, \Delta(\mathcal{F}^{(t-\epsilon)}))$. Choose any $f \in \mathcal{F}_\xi^{(t+\langle y', \xi \rangle)m} R_{m,\alpha} = \mathcal{F}^{(t+\langle y', \xi \rangle)m - \langle \alpha, \xi \rangle} R_{m,\alpha}$. Consider two cases:

  (i) $\langle \alpha/m, \xi \rangle - \langle y', \xi \rangle < \epsilon$. Then $\mathfrak{v}(f) \in \Delta^{(t-\epsilon)}$, so $|\mathfrak{v}(f)/m - y| \geq \delta_1$.
  (ii) $\langle \alpha/m, \xi \rangle - \langle y', \xi \rangle \geq \epsilon$. Then $|\mathfrak{v}(f)/m - y| \geq |\alpha/m - y'| \geq \epsilon/|\xi|$.

The two cases together imply that $y \notin \Delta(\mathcal{F}_\xi^{(t+\langle y', \xi \rangle)})$. So we get the inequality $G_{\mathcal{F}_\xi} \leq G_\mathcal{F} + \langle y', \xi \rangle$.

On the other hand, since $\mathcal{F} = (\mathcal{F}_\xi)_{-\xi}$, we also get $G_\mathcal{F} \leq G_{\mathcal{F}_\xi} - \langle y', \xi \rangle$. So we get the desired identity. $\square$

2.4 Asymptotically equivalent filtrations

In this section we recall Boucksom and Jonsson's characterization in [20; 21] of asymptotically equivalent filtrations; see also [1].

For a filtration $\mathcal{F} R_m$ of R_m, we say that a basis $\mathcal{B} = \{s_1, \ldots, s_{N_m}\}$ of R_m is compatible with $\mathcal{F} R_m$ if for any $\lambda \in \mathbb{R}$ there exists a subset of $\mathcal{B}$ that spans $\mathcal{F}^\lambda R_m$.

Let $\mathcal{F}_i = \{\mathcal{F}_i R_m\}$ for $i = 0, 1$ be two filtrations. For each m, we can find a basis $\mathcal{B} := \{s_1, \ldots, s_{N_m}\}$ of R_m that is compatible with both $\mathcal{F}_i R_m$ with $i = 0, 1$. We refer to [1; 18] and the discussion in Section 5 for more details. Assume that for each $i = 0, 1$, we have $s_k \in \mathcal{F}_i^{\mu_{k,i}} \setminus \mathcal{F}^{>\mu_{k,i}}$. Then $\mathcal{B}$ is an orthogonal basis for the non-Archimedean norm $\|\cdot\|_{m,i}$ corresponding to $\mathcal{F}_i$: for any $s = \sum_k a_k s_k \in R_m$,

$$(68) \qquad \|s\|_{m,i} = e^{-\max\{\lambda | s \in \mathcal{F}^\lambda R_m\}} = \max_k |a_k|_0 e^{-\mu_{k,i}},$$

where $|\cdot|_0$ is the trivial norm on $\mathbb{C}$.

Following [24; 20], we define the set of successive minima of $\mathcal{F}_1$ with respect to $\mathcal{F}_0$ to be the set $\{\mu_{k,1} - \mu_{k,0}\}$. The following result was proved in [18; 24].

Theorem 2.23 [18; 24]   *As $m \to +\infty$, the measures*

$$(69) \qquad \frac{n!}{m^n} \sum_{k=1}^{N_m} \delta_{(\mu_{k,1} - \mu_{k,0})/m}$$

converge weakly as $m \to +\infty$ to a relative limit measure, denoted by $d\nu := d\nu(\mathcal{F}_0, \mathcal{F}_1)$.

Corollary 2.24   *For any $p \in [1, \infty)$, the limit*

$$(70) \qquad d_p(\mathcal{F}_0, \mathcal{F}_1) := \lim_{m \to +\infty} \left(\frac{n!}{m^n} \sum_{k=1}^{N_m} m^{-1} |\mu_{k,1} - \mu_{k,0}|^p \right)^{1/p}$$

exists and is given by

$$(71) \qquad d_p(\mathcal{F}_0, \mathcal{F}_1) = \left(\int_{\mathbb{R}} |\lambda|^p \, d\nu(\lambda) \right)^{1/p}.$$

Definition 2.25 [20, Section 3.6]   *$\mathcal{F}_0$ and $\mathcal{F}_1$ are asymptotically equivalent if $d_2(\mathcal{F}_0, \mathcal{F}_1) = 0$.*

In fact, by [20] the d_p are comparable to each other for all $p \in [1, \infty)$, and the above equivalence can be defined by using any $p \in [1, +\infty)$.

Theorem 2.26 [20, Theorem 4.16]   *Assume that X is smooth. Let $\mathcal{F}_0$ and $\mathcal{F}_1$ be two filtrations on R. Then $\mathcal{F}_0$ and $\mathcal{F}_1$ are asymptotically equivalent if and only if $\phi_{\mathcal{F}_1} = \phi_{\mathcal{F}_2}$.*

We also need:

Proposition 2.27   *If $\mathcal{F}_{v_i}$ for $i = 0, 1$ are two $\mathbb{R}$–test configurations associated to two valuations $v_i \in \overset{\circ}{\mathrm{Val}}(X)$ for $i = 0, 1$, then $\phi_{\mathcal{F}_{v_1}} = \phi_{\mathcal{F}_{v_2}} + c$ for a constant $c \in \mathbb{R}$ if and only if $v_1 = v_2$ (and hence $\mathcal{F}_{v_1} = \mathcal{F}_{v_2}$).*

Proof Recall that $\phi_{\mathcal{F}_{v_i}} = \lim_{m \to +\infty} \phi_m^{\mathcal{F}_{v_i}}$ is an increasing limit along the subsequence $m = 2^k$, where for any $w \in \mathrm{Val}(X)$,

$$(72) \qquad \phi_m^{\mathcal{F}_{v_i}}(w) = -\frac{1}{m} G(w)\left(\sum_{\lambda \in \mathbb{N}} I_{m,\lambda}^{\mathcal{F}_{v_i}} t^{-\lambda} \right),$$

where $I_{m,\lambda}^{\mathcal{F}_{v_i}}$ is the base ideal of the sublinear system $\mathcal{F}_{v_i}^\lambda R_m$. Note that it is easy to see that $v_1 = v_2$ if and only if $\mathfrak{a}_\lambda(v_1) = \mathfrak{a}_\lambda(v_2)$ for any $\lambda \in \mathbb{N}$, where $\mathfrak{a}_\lambda(v_i) = \{f \in \mathcal{O}_X \mid v_i(f) \geq \lambda\}$.

For any $d \in \mathbb{N}$, by choosing $m \gg 1$ we can assume that $mL \otimes \mathfrak{a}_d(v_1)$ is globally generated. Then we get $I_{m,d}^{\mathcal{F}_{v_1}} = \mathfrak{a}_d(v_1)$. From this it is also clear that $\phi_{\mathcal{F}_{v_i}}(v_i) = 0$. So we get

$$-c = -\phi_{\mathcal{F}_{v_1}}(v_2) \leq -\phi_{2^k}^{\mathcal{F}_{v_1}}(v_2) = \frac{1}{2^k} G(v_2)\left(\sum_\lambda I_{2^k,\lambda}^{\mathcal{F}_{v_1}} t^{-\lambda} \right) \leq \frac{1}{2^k}\left(v_2(I_{2^k,d}^{\mathcal{F}_{v_1}}) - d \right) = \frac{1}{2^k}\left(v_2(\mathfrak{a}_d(v_1)) - d \right).$$

Since k can be arbitrarily large, we get $-c \leq 0$, ie $c \geq 0$. Switching v_1 and v_2 in the above argument, we get $c \leq 0$. So $c = 0$. We then have the inequality $v_2(\mathfrak{a}_d(v_1)) \geq d$ for any $d \in \mathbb{N}$. This easily implies $v_2 \geq v_1$. Switching v_1 and v_2, we get $v_1 \leq v_2$. Hence $v_1 = v_2$, as required. $\qquad\square$

Corollary 2.28 *Assume that X is smooth. With the same notation as above, if $\mathcal{F}_{v_1}$ is asymptotically equivalent to $\mathcal{F}_{v_2}$, then $v_1 = v_2$.*

More recently, this result has been proved for any $\mathbb{Q}$–Fano variety:

Lemma 2.29 [15, Lemma 3.16; 21, Theorem C] *For any $\mathbb{Q}$–Fano variety, if v_i for $i = 1, 2$ are two valuations in $\overset{\circ}{\mathrm{Val}}(X)$ such that $\mathcal{F}_{v_1}$ is asymptotically equivalent to $\mathcal{F}_{v_2}$, then $v_1 = v_2$.*

Remark 2.30 In the first version of this paper, Corollary 2.28 was stated for any $\mathbb{Q}$–Fano variety. However, it has been pointed out by experts that the validity of Theorem 2.26 from [20] for singular $\mathbb{Q}$–Fano varieties depends on a still conjectural property called *continuity of envelopes*. Fortunately, recently, in [15; 21], the result in Lemma 2.29 has been given a direct proof without using the continuity of envelopes.

2.5 Non-Archimedean invariants of filtrations

For any filtration $\mathcal{F}$ on $R = R(X, -K_X)$, we set

$$(73) \qquad \boldsymbol{L}^{\mathrm{NA}}(\phi^{\mathcal{F}}) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{L}_X^{\mathrm{NA}}(\mathcal{F}) = \inf_{v \in X_{\mathbb{Q}}^{\mathrm{div}}} \left(A_X(v) + (\phi_{\mathcal{F}} - \phi_{\mathrm{triv}})(v) \right),$$

$$(74) \qquad \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\phi^{\mathcal{F}}) = \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}) = \widetilde{\boldsymbol{S}}_X^{\mathrm{NA}}(\mathcal{F}) = -\log\left(\frac{1}{V} \int_{\mathbb{R}} e^{-\lambda} \, \mathrm{DH}(\mathcal{F}) \right) = -\log\left(\frac{n!}{V} \int_\Delta e^{-G_{\mathcal{F}}(y)} \, dy \right),$$

$$(75) \qquad \boldsymbol{E}^{\mathrm{NA}}(\phi^{\mathcal{F}}) = \boldsymbol{E}^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{E}_X^{\mathrm{NA}}(\mathcal{F}) = \frac{1}{V} \int_{\mathbb{R}} \lambda \cdot \mathrm{DH}(\mathcal{F}) = \frac{n!}{V} \int_\Delta G_{\mathcal{F}}(y) \, dy,$$

$$(76) \qquad \boldsymbol{H}^{\mathrm{NA}}(\phi^{\mathcal{F}}) = \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{H}_X^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}),$$

$$(77) \qquad \boldsymbol{D}^{\mathrm{NA}}(\phi^{\mathcal{F}}) = \boldsymbol{D}^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{D}_X^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}) - \boldsymbol{E}^{\mathrm{NA}}(\mathcal{F}).$$

The above functionals are by now well known, and we use notation following that in [19; 43]. The formula involving $G_{\mathcal{F}}$ follows from Theorem 2.5(ii).

Proposition 2.31 (see [37; 20; 52]) *For a filtration* $\mathcal{F}$, *with the notation from Definition 2.17, we have the following convergence*: *the sequence from Definition–Proposition 2.15 satisfies, for any* $\boldsymbol{F} \in \{\tilde{\boldsymbol{S}}, \boldsymbol{E}\}$,

$$\lim_{m\to+\infty} \boldsymbol{F}^{\mathrm{NA}}(\phi_m^{\mathcal{F}}) = \boldsymbol{F}^{\mathrm{NA}}(\phi^{\mathcal{F}}). \tag{78}$$

Moreover, we have

$$\lim_{m\to+\infty} \boldsymbol{L}^{\mathrm{NA}}(\phi_m^{\mathcal{F}}) \le \boldsymbol{L}^{\mathrm{NA}}(\phi^{\mathcal{F}}). \tag{79}$$

Proof By Proposition 2.16 we know that $\mathrm{DH}(\mathcal{F}_{(\mathcal{X}_m, \mathcal{L}_m, \eta_m)})$ converges weakly to $\mathrm{DH}(\mathcal{F})$ as $m \to +\infty$, from this we easily get the convergence of $\tilde{\boldsymbol{S}}^{\mathrm{NA}}$ and $\boldsymbol{E}^{\mathrm{NA}}$.

The inequality (79) follows easily from the inequality $\phi_m^{\mathcal{F}} \le \phi^{\mathcal{F}}$. $\qquad\square$

For our later argument, we will use a different formulation of the $\boldsymbol{L}^{\mathrm{NA}}$–functional studied in [80; 15]. For any filtration $\mathcal{F}$, denote by $I_\bullet^{\mathcal{F}^{(x)}} = \{I_{m,mx}^{\mathcal{F}}\}$ the graded sequence of base ideals defined in (52). In [79], Xu and Zhuang introduced the functional

$$\hat{\boldsymbol{L}}^{\mathrm{NA}}(\mathcal{F}) = \sup\{x \in \mathbb{R} \mid \mathrm{lct}(X; I_\bullet^{\mathcal{F}^{(x)}}) \ge 1\}, \tag{80}$$

and proved that $\hat{\boldsymbol{L}}^{\mathrm{NA}}(\mathcal{F}) \ge \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F})$. More recently it has been shown that in fact the two functionals are identical to each other. More specifically, we will need the following comparison results.

Proposition 2.32 [79, Proposition 4.2 and Theorem 4.3; 15, Lemma 3.8] *For any filtration* $\mathcal{F}$, *we have*:

(i) $A_X(E) \ge \hat{\boldsymbol{L}}^{\mathrm{NA}}(\mathcal{F}_{\mathrm{ord}_E})$ *for any prime divisor* E *over* X, *with equality holding if* ord_E *induces a weakly special test configuration.*

(ii) $\hat{\boldsymbol{L}}^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F})$ *for any filtration* $\mathcal{F}$.

For later purposes, we also introduce, for any $a > 0$,

$$\boldsymbol{E}_k^{\mathrm{NA}}(\mathcal{F}) = \frac{1}{V} \int_{\mathbb{R}} x^k \, \mathrm{DH}(\mathcal{F}) = \lim_{m\to+\infty} \frac{1}{N_m} \sum_i \left(\frac{\lambda_i^{(m)}}{m}\right)^k, \tag{81}$$

$$\boldsymbol{Q}^{(a)}(\mathcal{F}) = \frac{1}{V} \int_{\mathbb{R}} e^{-ax} \, \mathrm{DH}(\mathcal{F}) = \frac{1}{V} \sum_{k=0}^{+\infty} \frac{(-1)^k}{k!} a^k \boldsymbol{E}_k^{\mathrm{NA}}(\mathcal{F}), \tag{82}$$

$$\boldsymbol{Q}(\mathcal{F}) := \boldsymbol{Q}^{(1)}(\mathcal{F}). \tag{83}$$

Note that $\boldsymbol{E}_1^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{E}^{\mathrm{NA}}(\mathcal{F})$ and $\tilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}) = -\log \boldsymbol{Q}(\mathcal{F})$.

For any $v \in \overset{\circ}{\mathrm{Val}}(X)$ (resp. test configuration $(\mathcal{X}, \mathcal{L}, a\eta)$), we will often write $\boldsymbol{F}^{\mathrm{NA}}(v)$ (resp. $\boldsymbol{F}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta)$) for the above various functionals $\boldsymbol{F}^{\mathrm{NA}}(\mathcal{F})$ with $\mathcal{F}$ being the corresponding filtration.

Example 2.33 If $(\mathcal{X}, \mathcal{L}, a\eta)$ is a normal $\mathbb{R}$–test configuration, then we have

$$(84) \qquad \boldsymbol{E}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = a \cdot \frac{\overline{\mathcal{L}}^{\cdot n+1}}{(n+1)\boldsymbol{V}},$$

$$(85) \qquad \boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = a \cdot (\mathrm{lct}(\mathcal{X}, -K_{\mathcal{X}} - \mathcal{L}; \mathcal{X}_0) - 1).$$

If $(\mathcal{X}, \mathcal{X}_0)$ has log canonical singularities and $K_{\mathcal{X}} + \mathcal{L} = \sum_i e_i E_i$ (which is centered at $\mathcal{X}_0$), then

$$(86) \qquad \boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = a \cdot \min_i e_i.$$

Example 2.34 Let $\mathcal{F}$ be a special $\mathbb{R}$–test configuration and let $(X_0, \eta) = (X_{\mathcal{F},0}, \eta_{\mathcal{F}})$ be the corresponding central fiber. Assume that $\mathcal{F}|_{X_0} = \mathcal{F}'_{\mathrm{wt}_\eta} R'(-\sigma)$; see (33). Let $\widetilde{\varphi}$ be any $(S^1)^r$–invariant smooth positively curved Hermitian metric on $-K_X$. Then with the notation as in the paragraph containing (9), we have

$$(87) \qquad \boldsymbol{L}_X^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{L}_{X_0}^{\mathrm{NA}}(\mathcal{F}|_{X_0}) = \frac{\int_{X_0} \theta_{\widetilde{\varphi}}(\eta) e^{-\widetilde{\varphi}}}{\int_{X_0} e^{-\widetilde{\varphi}}} - \sigma = -\sigma,$$

$$(88) \qquad \boldsymbol{E}_X^{\mathrm{NA}}(\mathcal{F}) = \boldsymbol{E}_{X_0}^{\mathrm{NA}}(\mathcal{F}|_{X_0}) = \frac{1}{\boldsymbol{V}} \int_{X_0} \theta_{\widetilde{\varphi}}(\eta)(\mathrm{dd}^c\widetilde{\varphi})^n - \sigma,$$

$$(89) \qquad \widetilde{\boldsymbol{S}}_X^{\mathrm{NA}}(\mathcal{F}) = \widetilde{\boldsymbol{S}}_{X_0}^{\mathrm{NA}}(\mathcal{F}|_{X_0}) = -\log\left(\frac{1}{\boldsymbol{V}} \int_{X_0} e^{-\theta_\eta}(\mathrm{dd}^c\widetilde{\varphi})^n\right) - \sigma.$$

The above identity is well known if $\mathcal{F}$ comes from a special test configuration. For more general $\mathcal{F}$, one can use a sequence of special test configuration to approximate and get the above formula.

Corresponding to (58), we have the following simple transformation rule, which can be checked easily from the defining expressions of the functionals.

Lemma 2.35 *For any* $(a, b) \in \mathbb{R}_{>0} \times \mathbb{R}$, *we have*

$$(90) \qquad \boldsymbol{L}^{\mathrm{NA}}(a\mathcal{F}(b)) = \widehat{\boldsymbol{L}}^{\mathrm{NA}}(a\mathcal{F}(b)) = a\boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}) + b,$$

$$(91) \qquad \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}(b)) = \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}) + b,$$

$$(92) \qquad \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}(b)) = \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}).$$

We also note:

Lemma 2.36 *The function* $a \mapsto \boldsymbol{H}^{\mathrm{NA}}(a\mathcal{F})$ *is a convex function on* $\mathbb{R}_{\geq 0}$.

Proof Since $\boldsymbol{L}^{\mathrm{NA}}(a\mathcal{F})$ is linear in a by (90), we just need to show that $f(a) := -\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(a\mathcal{F})$ is convex in $a \in \mathbb{R}_{\geq 0}$. By (25) and (74), we get

$$f(a) = \log\left(\frac{n!}{\boldsymbol{V}} \int_\Delta e^{-aG(y)} \, dy\right),$$

where $G = G(y) = G_{\mathcal{F}}(y)$. So we can calculate

$$f''(a) = \frac{\int_\Delta G^2 e^{-aG}\, dy}{\int_\Delta e^{-aG}\, dy} - \frac{\left(\int_\Delta G e^{-aG}\, dy\right)^2}{\left(\int_\Delta e^{-aG}\, dy\right)^2} \geq 0$$

by Hölder's inequality.   □

Note the first identity in (90) comes from Proposition 2.32. Moreover, combining [52, Lemma 3.10] and Proposition 2.32, we have the following invariance property under the twisting.

Lemma 2.37 Let $\mathcal{F}$ be a $\mathbb{T}$–equivariant filtration. For any $\xi \in N_{\mathbb{R}}$, we have

$$(93) \qquad\qquad \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}_\xi) = \widehat{\boldsymbol{L}}^{\mathrm{NA}}(\mathcal{F}_\xi) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}).$$

As a consequence, we have

$$(94) \qquad\qquad \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}_{\mathrm{wt}_\xi}) = \widehat{\boldsymbol{L}}^{\mathrm{NA}}(\mathcal{F}_{\mathrm{wt}_\xi}) = 0.$$

The following lemma is a prototype uniqueness result in this paper, and can be seen as a generalization of the uniqueness of Kähler–Ricci soliton vector fields shown by Tian and Zhu [71] (the case when $\mathcal{F} = \mathcal{F}_{\mathrm{triv}}$). See Section 2.6 for more discussion.

Lemma 2.38 Let $\mathcal{F}$ be a $\mathbb{T}$–equivariant filtration. Then the function $\xi \mapsto \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}_\xi)$ on $N_{\mathbb{R}}$ admits a unique minimizer.

Proof By (93), $\boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}_\xi)$ is constant in ξ. Using the identity (67) and (74),

$$-\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}_\xi) = \log\left(\frac{n!}{V}\int_\Delta e^{-G_{\mathcal{F}_\xi}(y)}\, dy\right) = \log\left(\frac{n!}{V}\int_\Delta e^{-G_{\mathcal{F}}(y)-\langle y,\xi\rangle}\, dy\right).$$

It is easy to use this expression to show that $f(\xi) := -\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}_\xi)$ is strictly convex in $\xi \in N_{\mathbb{R}}$, which implies the uniqueness of minimizer. To prove the existence of minimizer, we need to show that $f(\xi)$ is proper, ie $\lim_{|\xi|\to+\infty} f(\xi) = +\infty$. To see this, recall that we have the vanishing

$$(95) \qquad\qquad \int_X \theta_{\widetilde{\varphi}}(\xi) e^{-\widetilde{\varphi}} = -\int_X \mathfrak{L}_\eta e^{-\widetilde{\varphi}} = 0.$$

This implies that $0 > \inf_X \theta_{\widetilde{\varphi}}(\xi) = \inf_\Delta \langle y,\xi\rangle$ if $\xi \neq 0$, which indeed implies the properness.   □

Definition 2.39 We say that a filtration $\mathcal{F}$ is normalized if

$$(96) \qquad\qquad \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}) = 0.$$

A test configuration $(\mathcal{X}, \mathcal{L}, a\eta)$ is normalized if $\mathcal{F}_{(\mathcal{X},\mathcal{L},a\eta)}$ is normalized.

With the above discussion, the following lemma is easy to prove.

Lemma 2.40   (i) *Any special test configuration $(\mathcal{X}, -K_{\mathcal{X}})$ is normalized. More generally, a special $\mathbb{R}$–test configuration $\mathcal{F}$ (see Definition 2.8) if and only if $\sigma = 0$ in (33).*

   (ii) *For any filtration $\mathcal{F}$, the shift $\mathcal{F}(-\boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}))$ is normalized. If $\mathcal{F}$ is normalized, then so are $a\mathcal{F}$ for any $a > 0$, and any twist $\mathcal{F}_{\xi}$.*

As a consequence of this approximation result in Proposition 2.31, it is convenient for us to introduce:

Definition–Proposition 2.41   *For any $\mathbb{Q}$–Fano variety X, we define*

$$\tag{97} h(X) = \inf_{(\mathcal{X}, \mathcal{L}, a\eta)} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = \inf_{\mathcal{F}} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}),$$

where $(\mathcal{X}, \mathcal{L}, a\eta)$ ranges over all test configurations, and $\mathcal{F}$ ranges over all filtrations or $\mathbb{R}$–test configurations.

The following lemma is similar to [31, Lemma 2.5].

Lemma 2.42   *For any filtration $\mathcal{F}$, we have*

$$\tag{98} \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}) \leq \boldsymbol{E}^{\mathrm{NA}}(\mathcal{F}) \quad \text{and} \quad \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}) \geq \boldsymbol{D}^{\mathrm{NA}}(\mathcal{F}).$$

The identities hold true if and only if $\mathcal{F}(c)$ is asymptotically equivalent to the trivial filtration for some $c \in \mathbb{R}$; see Definition 2.25.

Proof   The first inequality, which implies the second, follows from the concavity of the logarithmic function. When the identity holds, the DH measure $\mathrm{DH}(\mathcal{F})$ is a Dirac measure $V \cdot \delta_c$. Then $d_2(\mathcal{F}(c), \mathcal{F}_{\mathrm{triv}}) = 0$, which by Definition 2.25 means that $\mathcal{F}(c)$ is asymptotically equivalent to the trivial filtration.   □

Based on the work in [29; 28; 40], Dervan and Székelyhidi proved:

Theorem 2.43   [31]   *Assume that X is a smooth Fano manifold. There is an identity*

$$\tag{99} h(X) := \inf_{(\mathcal{X}, \mathcal{L}, a\eta) \text{ special}} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = - \inf_{\omega \in c_1(X)} \int_X h_{\omega} e^{h_{\omega}} \omega^n,$$

where ω ranges over smooth Kähler metrics from $c_1(X)$ and h_{ω} is the normalized Ricci potential of ω. Moreover, the infimum is achieved by a special test configuration constructed via the Gromov–Hausdorff limit Kähler–Ricci soliton from [29; 28].

More recently, Hisamoto [43] gave a different proof of (99) based on the destabilizing geodesic rays constructed from [30].

Remark 2.44   Our sign convention differs from that of Dervan–Székelyhidi and Hisamoto by a minus. Dervan and Székelyhidi defined a non-Archimedean functional for general $\mathbb{R}$–test configuration by mimicking Tian's CM weight (or the so-called Donaldson–Futaki invariant). But in such generality, their

normalization seems imprecise. Differently from their definition, for any test configuration $(\mathcal{X}, \mathcal{L}, a\eta)$ one could define

$$\tilde{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = \frac{a}{V}(K_{\overline{\mathcal{X}}/\mathbb{P}^1} \cdot \overline{\mathcal{L}}^{\cdot n} + \overline{\mathcal{L}}^{\cdot n+1}) - \tilde{S}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta). \tag{100}$$

By the same argument as [19, Proposition 7.32], we have

$$H^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, \eta) \leq \tilde{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, \eta), \tag{101}$$

with strict inequality if $(\mathcal{X}, \mathcal{L}, \eta)$ is anticanonical. Moreover, by (98) we also get

$$\tilde{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, \eta) \geq \mathrm{CM}(\mathcal{X}, \mathcal{L}, \eta) = \frac{1}{V}\left(K_{\overline{\mathcal{X}}/\mathbb{P}^1} \cdot \overline{\mathcal{L}}^{\cdot n} + \frac{n}{n+1}\overline{\mathcal{L}}^{\cdot n+1}\right), \tag{102}$$

with the identity being true only if $(\mathcal{X}, \mathcal{L}, \eta)$ is trivial. One advantage of H^{NA} over $\tilde{H}^{\mathrm{NA}}$ is that the former can be defined for any filtration, not necessarily finitely generated. Due to this reason, we will not use $\tilde{H}^{\mathrm{NA}}$ in this paper.

2.6 g–Ding-stability and Kähler–Ricci solitons

Let $\mathcal{F}$ be a $\mathbb{T}$–equivariant filtration. For any $\lambda \in \mathbb{R}$, we have a (finite) decomposition

$$\mathcal{F}^\lambda R_m = \bigoplus_{\alpha \in M_{\mathbb{Z}}} \mathcal{F}^\lambda R_{m,\alpha}. \tag{103}$$

Let P be the moment polytope of $(X, -K_X)$ with respect to the $\mathbb{T}$–action. Let g be a smooth positive function on P. Fix a faithful $\mathbb{Z}^n$–valuation that is adapted to the torus action (see Definition 2.21) and let $\Delta \subset \mathbb{R}^n$ be the Okounkov body that satisfies (65): $p(\Delta) = P$, where $p \colon \mathbb{R}^n \to \mathbb{R}^r$ is the natural projection. Still denote by g the function p^*g on Δ. Define the g–volume of graded linear series $\{\mathcal{F}^{(t)} R_m\}$ as

$$\mathrm{vol}_g(\mathcal{F}^{(t)}) := \lim_{m \to +\infty} \sum_\alpha g\left(\frac{\alpha}{m}\right)\frac{\dim \mathcal{F}^{mt} R_{m,\alpha}}{m^n/n!} = n! \cdot \int_{\Delta(\mathcal{F}^{(t)})} g(y) \, dy_{\mathrm{Leb}} =: n! \cdot \mathrm{vol}_g(\Delta(\mathcal{F}^{(t)})).$$

Then, as in the $g \equiv 1$ case, we have the convergence

$$\mathrm{DH}_g(\mathcal{F}) := \lim_{m \to +\infty} \sum_\alpha g\left(\frac{\alpha}{m}\right)\delta_{\lambda_i^{(m,\alpha)}/m} = -d\,\mathrm{vol}_g(\mathcal{F}^{(t)}) = n! \cdot (G_{\mathcal{F}})_*(g(y)\, dy_{\mathrm{Leb}}).$$

We also set

$$V_g := n! \cdot \mathrm{vol}_g(\Delta) = n! \cdot \int_\Delta g(y) \, dy_{\mathrm{Leb}} = \int_{\mathbb{R}} \mathrm{DH}_g(\mathcal{F}), \tag{104}$$

$$E_g^{\mathrm{NA}}(\mathcal{F}) := \frac{n!}{V_g} \int_\Delta G_{\mathcal{F}}(y)g(y)\, dy_{\mathrm{Leb}} = \frac{1}{V_g} \int_{\mathbb{R}} \lambda \cdot \mathrm{DH}_g(\mathcal{F}), \tag{105}$$

$$D_g^{\mathrm{NA}}(\mathcal{F}) := L^{\mathrm{NA}}(\mathcal{F}) - E_g^{\mathrm{NA}}(\mathcal{F}). \tag{106}$$

If $(\mathcal{X}, \mathcal{L}, \eta)$ is a test configuration, then we set $D_g^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, \eta) = D_g^{\mathrm{NA}}(\mathcal{F}_{(\mathcal{X}, \mathcal{L}, \eta)})$.

Definition 2.45 (X, ξ) is g–Ding-semistable if $\boldsymbol{D}_g^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, \eta) \geq 0$ for any $\mathbb{T}$–equivariant test configuration $(\mathcal{X}, \mathcal{L}, \eta)$ of $(X, -K_X)$.

(X, ξ) is g–Ding-polystable if it is g–Ding-semistable, and $\boldsymbol{D}_g^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, \eta) = 0$ for a $\mathbb{T}$–equivariant weakly special test configuration (see Definition 2.12) only if $(\mathcal{X}, \mathcal{L}, \eta)$ is a product test configuration.

The following result was proved by adapting the techniques of MMP from [56; 37; 7].

Theorem 2.46 (see [39]) *To test the g–Ding-semistability, or the g–Ding-polystability, of (X, ξ), it suffices to test over all special test configurations.*

We have the following valuative criterion:

Theorem 2.47 [39] *X is g–Ding-semistable if and only if for any $v \in (X_{\mathbb{Q}}^{\mathrm{div}})^{\mathbb{T}}$, we have*

$$\text{(107)} \qquad \beta_g(v) := A_X(v) - \frac{1}{V_g} \int_0^{+\infty} \mathrm{vol}_g(\mathcal{F}_v^{(t)}) \, dt \geq 0.$$

Now we use our notation to reformulate the holomorphic invariants of Tian and Zhu [71] in the study of Kähler–Ricci solitons. We refer to [71; 8; 39] for more details and references. Let X be a $\mathbb{Q}$–Fano variety with an effective $\mathbb{T}$–action. We use the same notation, such as an $(S^1)^r$–invariant smooth Hermitian metric $\widetilde{\varphi}$ on $-K_X$, a moment polytope $P \subset M_{\mathbb{R}}$, a function $\theta_{\widetilde{\varphi}}(\eta) = \mathcal{L}_\eta e^{-\widetilde{\varphi}}/e^{-\widetilde{\varphi}}$, etc. We identify any $\eta \in N_{\mathbb{R}}$ with the corresponding holomorphic vector field on X.

A Kähler–Ricci soliton on (X, ξ) is a positively curved bounded Hermitian metric $e^{-\varphi}$ on $-K_X$ that satisfies the equation

$$\text{(108)} \qquad e^\varphi (\mathrm{dd}^c \varphi)^n = e^{\theta_\varphi(\xi)},$$

where $\theta_\varphi(\xi) = \theta_{\widetilde{\varphi}}(\xi) + \xi(\varphi - \widetilde{\varphi})$. Over X^{reg}, φ is smooth [6; 39] and satisfies the identity

$$\text{(109)} \qquad \mathrm{Ric}(\mathrm{dd}^c \varphi) - \mathrm{dd}^c \varphi = -\mathrm{dd}^c \theta_\varphi(\xi).$$

As a consequence, the family of metrics $\varphi(s) := \sigma_\xi(s)^* \varphi$ satisfies the normalized Kähler–Ricci flow,

$$\text{(110)} \qquad \frac{d}{ds} \mathrm{dd}^c \varphi(s) = -\mathrm{Ric}(\mathrm{dd}^c \varphi(s)) + \mathrm{dd}^c \varphi(s).$$

For any $\xi \in N_{\mathbb{R}}$, we set $g_\xi(x) = e^{-\langle x, \xi \rangle} = e^{-\sum_{i=1}^r \xi^i x_i}$, which is a smooth positive function on P, and write $\boldsymbol{F}_{g_\xi}$ as $\boldsymbol{F}_\xi$ for $\boldsymbol{F} \in \{\boldsymbol{L}, \boldsymbol{D}\}$ etc., and $V_\xi := V_{g_\xi}$. Tian and Zhu [71] defined a modified Futaki invariant as an obstruction to the existence of Kähler–Ricci solitons on (X, ξ): for any $\eta \in N_{\mathbb{R}}$,

$$\text{(111)} \qquad \mathrm{Fut}_\xi(\eta) := -\frac{1}{V_\xi} \int_X \theta_{\widetilde{\varphi}}(\eta) e^{-\theta_{\widetilde{\varphi}}(\xi)} (\mathrm{dd}^c \widetilde{\varphi})^n = \boldsymbol{D}_\xi^{\mathrm{NA}}(\mathrm{wt}_\eta),$$

where $V_\xi = \int_X e^{-\theta_{\widetilde{\varphi}}(\xi)} (\mathrm{dd}^c \widetilde{\varphi})^n$. The second identity follows by noting that $\boldsymbol{D}_\xi^{\mathrm{NA}}(\mathrm{wt}_\eta) = -\boldsymbol{E}_\xi^{\mathrm{NA}}(\mathrm{wt}_\eta)$ because of the vanishing $\boldsymbol{L}^{\mathrm{NA}}(\mathrm{wt}_\eta) = 0$.

Remark 2.48 Again, here we have used the negative sign convention compared to [71].

Fut$_\xi$ does not depend on the choice of $\widetilde{\varphi}$ and (X, ξ) admits a KR soliton only if Fut$_\xi \equiv 0$ on $N_{\mathbb{R}}$. Moreover, by [71, Lemma 2.2] the soliton vector field is a priori uniquely determined by minimizing the strictly convex functional (Tian and Zhu didn't use the logarithm) on $N_{\mathbb{R}}$ (see Lemma 2.38), which is the antiderivative of $\eta \mapsto$ Fut$_\xi(\eta)$,

$$(112) \qquad \xi \mapsto \log\left(\frac{1}{V}\int_X e^{-\theta_{\widetilde{\varphi}}(\xi)}(\mathrm{dd}^c\widetilde{\varphi})^n\right) = \log\left(\frac{1}{V}\int_{\mathbb{R}} e^{-\lambda}\,\mathrm{DH}(\mathcal{F}_{\mathrm{wt}_\xi})\right) = -\widetilde{S}^{\mathrm{NA}}(\mathrm{wt}_\xi).$$

Recall also that $L^{\mathrm{NA}}(\mathrm{wt}_\eta) = \widehat{L}^{\mathrm{NA}}(\mathrm{wt}_\eta) \equiv 0$ on $N_{\mathbb{R}}$; see (94). Combining these discussions we get the derivative identity

$$(113) \qquad \frac{d}{ds}H^{\mathrm{NA}}(\mathrm{wt}_{\xi+s\eta}) = \frac{d}{ds}H^{\mathrm{NA}}(\mathrm{wt}_{\xi+s\eta}) = D_\xi^{\mathrm{NA}}(\mathrm{wt}_\eta) = \mathrm{Fut}_\xi(\eta).$$

For simplicity of notation, we introduce:

Definition 2.49 We say that (X, ξ) is K–semistable (resp. K–polystable) if X is g_ξ–Ding-semistable (resp. g_ξ–Ding-polystable).

Remark 2.50 Since by Theorem 2.46 it is enough to test the stability on special test configurations, this definition coincides with the original modified K–(poly)stability adopted by Tian as well as Berman, Witt and Nyström, and others. To respect the original notation, we will just call (X, ξ) K–(poly)stable, although we will also freely use the notion of Ding-(poly)stability.

By [8; 31], when X is smooth, the Yau–Tian–Donaldson conjecture is true, ie K–polystability is equivalent to the existence of Kähler–Ricci solitons. For singular X, we proved in [39] a version of the Yau–Tian–Donaldson conjecture involving $\mathrm{Aut}(X, \xi)_0$–uniform Ding-stability.

3 H^{NA} invariant and MMP

3.1 An intersection formula for higher moments

Let $(\mathcal{X}, \mathcal{L}, \eta)$ be any normal ample test configuration. Choose a smooth (semipositive) curvature form ω in $c_1(\mathcal{L}|_{\mathcal{X}_0})$. Let θ be the Hamiltonian function for η with respect to ω, so $\iota_\eta\omega = (\sqrt{-1}/2\pi)\,\overline{\partial}\theta$. By the equivariant Riemann–Roch formula, we get

$$E_k^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}) := E_k^{\mathrm{NA}}(\mathcal{F}_{(\mathcal{X}, \mathcal{L})}) = \lim_{m \to +\infty}\frac{1}{N_m}\sum_i\left(\frac{\lambda_i^{(m)}}{m}\right)^k = \frac{1}{V}\int_{\mathcal{X}_0}\theta^k\omega^n.$$

To motivate our calculations, we will first give a direct proof of two identities which can already be derived from the above discussion.

Lemma 3.1 *We have*

$$(114) \qquad E_k^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}) = \frac{1}{V}\int_{\mathbb{R}} x^k\,\mathrm{DH}(\mathcal{F}^{(x)}),$$

$$(115) \qquad E^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}) = E_1^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}) = \frac{1}{V}\frac{\overline{\mathcal{L}}^{\cdot n+1}}{n+1}.$$

Proof When we change $\mathcal{L}$ to $\mathcal{L} + d\mathcal{X}_0$, $\mathcal{F}$ is changed to $\mathcal{F}(d)$, and both sides of the above identities have d added to them. So we can assume that $\bar{\mathcal{L}}$ is very ample over $\bar{\mathcal{X}}$. Then we have

$$(116) \qquad \bar{\mathcal{X}} = \mathrm{Proj}\bigg(\bigoplus_{m \geq 0} \bigoplus_{j=0}^{+\infty} t^{-j} \mathcal{F}^j R_m \bigg)$$

and $\bar{\mathcal{L}}_d = \mathcal{O}_{\bar{\mathcal{X}}}(1)$.

For simplicity of notation, we write

$$(117) \qquad f_k(m) = \sum_{i=1}^{N_m} (\lambda_i^{(m)})^k = \sum_{j=0} j^k (\dim \mathcal{F}^j R_m - \dim \mathcal{F}^{j+1} R_m)$$
$$= \sum_{j=1} (j^k - (j-1)^k) \dim \mathcal{F}^j R_m = \sum_{j=1} (kj^{k-1} + O(j^{k-2})) \dim \mathcal{F}^j R_m.$$

We easily get the identity

$$(118) \quad E_k^{\mathrm{NA}} = \frac{1}{V} \lim_{m \to +\infty} \frac{n!}{m^{n+k}} f_k(m) = \frac{1}{V} \int_0^\infty k x^{k-1} \mathrm{vol}(\mathcal{F}^{(x)} R_\bullet) \, dx = \frac{1}{V} \int_{\mathbb{R}} x^k (-d\,\mathrm{vol}(\mathcal{F}^{(x)})).$$

Moreover, we have the dimension formula

$$\mathcal{N}_m := h^0(\bar{\mathcal{X}}, m\bar{\mathcal{L}}) = \sum_{j=0}^{+\infty} \dim \mathcal{F}^j R_m = \frac{m^{n+1}}{n!} \int_0^{+\infty} \mathrm{vol}(\mathcal{F}^{(x)} R_\bullet) \, dx + O(m^n),$$

which, by the Riemann–Roch formula, gives the identity

$$(119) \qquad \frac{1}{V} \frac{\bar{\mathcal{L}}^{\cdot n+1}}{n+1} = \frac{1}{V} \int_0^{+\infty} \mathrm{vol}(\mathcal{F}^{(x)} R_\bullet) \, dx = \frac{1}{V} \int_0^{+\infty} x \, \mathrm{DH}(\mathcal{F}). \qquad \square$$

The formula (115) goes back to Mumford's study of GIT [61], and has also been used in the study of K–stability. The following result is a generalization of it to higher moments. We will use the following notation as in [39]. Let $\mathbb{C}^* \to \mathbb{C}^{k+1} \setminus \{0\} \to \mathbb{P}^k$ be the principal $\mathbb{C}^*$–bundle and set

$$(120) \qquad (\bar{\mathcal{X}}^{[k]}, \bar{\mathcal{L}}^{[k]}) := ((\bar{\mathcal{X}}, \bar{\mathcal{L}}) \times (\mathbb{C}^{k+1} \setminus \{0\}))/\mathbb{C}^*.$$

Remark 3.2 Since the $\mathbb{C}^*$–action on $\bar{\mathcal{X}}$ moves the fiber $\bar{\mathcal{X}} \to \mathbb{P}^1$, the situation here is different from the situation in [19, Corollary 3.4] or [39], where a similar fiber construction with respect to a vertical torus action is used.

Proposition 3.3 *Let $(\mathcal{X}, \mathcal{L})$ be a normal ample test configuration. For any $k \geq 1$, we have the intersection formula*

$$(121) \qquad E_k^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}) = \frac{1}{V} \frac{k! n!}{(n+1)!} (\bar{\mathcal{L}}^{[k-1]})^{\cdot n+k}.$$

Proof We use the notation from the above proof and without loss of generality assume that $\bar{\mathcal{L}}$ is very ample over $\bar{\mathcal{X}}$.

The weights $\{\mu_\alpha \mid \alpha = 1, \ldots, \mathcal{N}_m\}$ and multiplicities of $\mathbb{C}^*$–action on $H^0(\bar{\mathcal{X}}, \bar{\mathcal{L}})$ are given according to the isomorphism (44). By the identity (117), the weight of $\mathbb{C}^*$ on $\det H^0(\bar{\mathcal{X}}, m\mathcal{L})$ is given by

$$(122) \qquad \sum_{\alpha=1}^{\mathcal{N}_m} \mu_\alpha^{k-1} = \sum_{j=0}^{+\infty} j^{k-1} \dim \mathcal{F}^j R_m = k^{-1} f_k(m) + O(m^{n+k-1}).$$

Choose a smooth Kähler metric $\Omega \in c_1(\bar{\mathcal{L}})$ on $\bar{\mathcal{X}}$ and let Θ be the Hamiltonian function for η. Then by the equivariant Riemann–Roch formula, we get

$$(123) \qquad \lim_{m \to +\infty} \frac{(n+1)!}{m^{n+1}} \sum_\alpha \left(\frac{\mu_\alpha}{m}\right)^{k-1} = \int_{\bar{\mathcal{X}}} \Theta^{k-1}\Omega^{n+1} = \frac{(k-1)!(n+1)!}{(k+n)!} \int_{\bar{\mathcal{X}}^{[k-1]}} (\Omega + \Theta t)^{n+k}$$

$$= \frac{(k-1)!(n+1)!}{(k+n)!} (\bar{\mathcal{L}}^{[k-1]})^{\cdot n+k}.$$

Combining (118), (122) and (123), we get

$$E_k^{\mathrm{NA}} = \frac{1}{V} \lim_{k \to} \frac{n!}{m^{n+k}} k \sum_\alpha \mu_\alpha^{k-1} = \frac{1}{V} \frac{k}{n+1} \frac{(k-1)!(n+1)!}{(k+n)!} (\bar{\mathcal{L}}^{[k-1]})^{\cdot n+k}$$

$$= \frac{1}{V} \frac{k!n!}{(k+n)!} (\bar{\mathcal{L}}^{[k-1]})^{\cdot n+k}. \qquad \qquad \square$$

Recall from (82) that $\widetilde{S}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = -\log Q^{(a)}$, where

$$Q^{(a)} = \frac{1}{V} \int_{\mathcal{X}_0} e^{-a\theta}\omega^n = \sum_{k=0}^\infty (-1)^k a^k \frac{1}{V} \int_{\mathcal{X}_0} \frac{\theta^k}{k!}\omega^n = \sum_k (-1)^k \frac{a^k}{k!} E_k^{\mathrm{NA}}.$$

Proposition 3.4 *Let* $(\mathcal{X}, \mathcal{L}_\lambda, a\eta)_{\lambda \in (-\epsilon, \epsilon)}$ *be a family of normal test configurations of* $(X, -K_X)$, *with a fixed total space and varying polarization. Assume that* $\mathcal{X}_0 = \sum_i b_i E_i$ *for irreducible components* E_i, *and that* $\mathcal{L}_\lambda$ *is differentiable with respect to* λ. *Then we have the derivative formula*

$$(124) \qquad \frac{d}{d\lambda} \widetilde{S}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) = a \frac{\sum_i e_i Q_i^{(a)}}{Q^{(a)}},$$

where $Q_i^{(a)} = (1/V) \int_{E_i} e^{-a\theta}\omega^n$.

Proof We use the intersection formula (121) to get

$$V \cdot \frac{d}{d\lambda} E_k^{\mathrm{NA}} = \frac{d}{d\lambda} \frac{k!n!}{(k+n)!} (\bar{\mathcal{L}}^{[k-1]})^{\cdot n+k} = \frac{k!n!}{(k+n-1)!} (\bar{\mathcal{L}}^{[k-1]})^{\cdot n+k-1} \cdot \dot{\bar{\mathcal{L}}}^{[k-1]}$$

$$= \frac{k!n!}{(k+n-1)!} \sum_i e_i \int_{E_i^{[k-1]}} (\Omega + \Theta t)^{n+k-1} = k \sum_i e_i \int_{E_i} \theta^{k-1}\omega^n,$$

where $E_i^{[k-1]} = (E_i \times \mathbb{C}^{k-1} \setminus \{0\})/\mathbb{C}^*$. So we get the desired formula

$$
\frac{d}{d\lambda} \boldsymbol{Q}^{(a)} = \sum_k (-1)^k \frac{a^k}{k!} \frac{d}{d\lambda} \boldsymbol{E}_k^{\mathrm{NA}} = \sum_{k=1} (-1)^k \frac{a^k}{(k-1)!} \sum_i e_i \frac{1}{V} \int_{E_i} \theta^{k-1} \omega^n
$$

$$
= -a \sum_i e_i \frac{1}{V} \int_{E_i} \sum_{j=0} \frac{(-1)^j}{j!} (a\theta)^j \omega^n = -a \sum_i e_i \frac{1}{V} \int_{E_i} e^{-a\theta} \omega^n = -a \sum_i \boldsymbol{Q}_i^{(a)}.
$$

The termwise differentiation and the change of summation are valid because of absolute convergence. $\square$

3.2 Decreasing of H^{NA} along MMP

Theorem 3.5 *Let $\mathbb{G}$ be a reductive group and $(\mathcal{X}, \mathcal{L}, a\eta)$ be a $\mathbb{G}$–equivariant normal test configuration. There exists a $\mathbb{G}$–equivariant special test configuration $(\mathcal{X}^s, \mathcal{L}^s, a^s\eta^s)$ such that*

$$
\text{(125)} \qquad\qquad H^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) \geq H^{\mathrm{NA}}(\mathcal{X}^s, \mathcal{L}^s, a^s\eta^s).
$$

Moreover, if $\mathcal{X}_0$ is reduced, then the identity holds true if $\mathcal{X}$ is already a special test configuration.

Proof For simplicity of notation, we assume $\mathbb{G}$ is trivial. The general case is obtained by running the $\mathbb{G}$–equivariant MMP in the following arguments.

Step 1 Choose a semistable reduction of $\mathcal{X} \to \mathbb{C}$. By this, we mean that there is an integer d and a $\mathbb{G}$–equivariant log resolution of singularities $\widetilde{\mathcal{X}} \to \mathcal{X}^{(d_1)} := (\mathcal{X} \times_{\mathbb{C}, t \to t^{d_1}} \mathbb{C})^{\mathrm{norm}}$ (see (49)) such that $(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}_0)$ is simple normal crossing. In particular, $\mathcal{X}_0^{(d_1)}$ is reduced. By using the identity (51) and Lemma 2.35 we easily get

$$
\text{(126)} \qquad\qquad H^{\mathrm{NA}}(\mathcal{X}^{(d_1)}, \mathcal{L}^{(d_1)}, a\eta^{(d_1)}/d_1) = H^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta).
$$

Step 2 In this step, we show that there exist $d_1 \in \mathbb{Z}_{>0}$, a projective birational $\mathbb{C}^*$–equivariant morphism $\pi \colon \mathcal{X}^{\mathrm{lc}} \to \mathcal{X}^{(d_1)}$ and a normal, ample test configuration $(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}^{\mathrm{lc}})/\mathbb{C}$ for (X, L), such that

$$
\text{(127)} \qquad\qquad H^{\mathrm{NA}}(\mathcal{X}^{(d_1)}, \mathcal{L}^{(d_1)}, a\eta^{(d_1)}/d_1) \geq H^{\mathrm{NA}}(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}^{\mathrm{lc}}, a\eta^{\mathrm{lc}}/d_1).
$$

Moreover, if the equality holds, then $(\mathcal{X}^{(d_1)}, \mathcal{L}^{(d_1)})$ is isomorphic to $(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}^{\mathrm{lc}})$, and hence $(\mathcal{X}, \mathcal{X}_0)$ is already log canonical.

We run a $\mathbb{C}^*$–equivariant MMP to get a log canonical modification $\pi^{\mathrm{lc}} \colon \mathcal{X}^{\mathrm{lc}} \to \mathcal{X}^{(d_1)}$ such that $(\mathcal{X}^{\mathrm{lc}}, \mathcal{X}_0^{\mathrm{lc}})$ is log canonical and $K_{\mathcal{X}^{\mathrm{lc}}}$ is relatively ample over $\mathcal{X}^{(d_1)}$. Set $E = K_{\mathcal{X}^{\mathrm{lc}}} + (\pi^{\mathrm{lc}})^*\mathcal{L} = \sum_{i=1}^k e_i \mathcal{X}_{0,i}$ with $e_1 \leq e_2 \leq \cdots \leq e_k$ and $\mathcal{L}_\lambda^{\mathrm{lc}} = (\pi^{\mathrm{lc}})^*\mathcal{L}^{(d_1)} + \lambda E$. Then since E is relatively ample over $\mathcal{X}^{(d_1)}$, $\mathcal{L}_\lambda$ is ample over $\mathcal{X}^{\mathrm{lc}}$ for $0 < \lambda \ll 1$. So

$$
\boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}_\lambda^{\mathrm{lc}}, a\eta^{\mathrm{lc}}/d_1) = \frac{a}{d_1} \boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}_\lambda^{\mathrm{lc}}, \eta^{\mathrm{lc}}) = \frac{a}{d_1}(1+\lambda)e_1.
$$

By definition (76), we have

$$
\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}_\lambda^{\mathrm{lc}}, a\eta^{\mathrm{lc}}/d_1) = -\log \boldsymbol{Q}^{(ad_1^{-1})}, \quad \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}_\lambda^{\mathrm{lc}}, a\eta^{\mathrm{lc}}/d_1) = \frac{a(1+\lambda)e_1}{d_1} + \log \boldsymbol{Q}^{(ad_1^{-1})}.
$$

We then use (124) to calculate

$$\frac{d}{d\lambda} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}^{\mathrm{lc}}_\lambda, a\eta^{\mathrm{lc}}/d_1) = \frac{ae_1}{d_1} - \frac{a}{d_1} \frac{\sum_i e_i \boldsymbol{Q}_i^{(ad_1^{-1})}}{\sum_i \boldsymbol{Q}_i^{(ad_1^{-1})}} = -\frac{a}{d_1} \frac{\sum_i (e_i - e_1) \boldsymbol{Q}_i^{(ad_1^{-1})}}{\boldsymbol{Q}^{(ad_1^{-1})}} \le 0.$$

The last identity holds if and only if $e_i \equiv e_1$, and hence $(\mathcal{X}^{(d_1)}, \mathcal{L}^{(d_1)}) \cong (\mathcal{X}^{\mathrm{lc}}, \mathcal{L}^{\mathrm{lc}})$. In this case, $(\mathcal{X}^{(d_1)}, \mathcal{X}_0^{(d_1)})$ is log canonical, which implies that $(\mathcal{X}, \mathcal{X}_0)$ is already log canonical, by the pullback formula for the log differential; see [56, page 210].

Step 3 With the $(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}^{\mathrm{lc}})$ obtained from the first step, we run a relative MMP with scaling to get a normal, ample test configuration $(\mathcal{X}^{\mathrm{ac}}, \mathcal{L}^{\mathrm{ac}})/\mathbb{P}^1$ for $(X, -K_X)$ with $(\mathcal{X}^{\mathrm{ac}}, \mathcal{X}_0^{\mathrm{ac}})$ log canonical such that $-K_{\mathcal{X}^{\mathrm{ac}}} \sim_{\mathbb{Q},\mathbb{C}} \mathcal{L}^{\mathrm{ac}}$. More concretely, we take $q \gg 1$ such that $\mathcal{H}^{\mathrm{lc}} = \mathcal{L}^{\mathrm{lc}} - (q+1)^{-1}(\mathcal{L}^{\mathrm{lc}} + K_{\mathcal{X}^{\mathrm{lc}}})$ is relatively ample. Set $\mathcal{X}^0 = \mathcal{X}^{\mathrm{lc}}$, $\mathcal{L}^0 = \mathcal{L}^{\mathrm{lc}}$, $\mathcal{H}^0 = \mathcal{H}^{\mathrm{lc}}$ and $\lambda_0 = q + 1$. Then $K_{\mathcal{X}^0} + \lambda_0 \mathcal{H}^0 = q\mathcal{L}^0$. We run a sequence of $K_{\mathcal{X}^0}$–MMP over $\mathbb{C}$ with scaling of $\mathcal{H}^0$. Then we obtain a sequence of models

$$\mathcal{X}^0 \dashrightarrow \mathcal{X}^1 \dashrightarrow \cdots \dashrightarrow \mathcal{X}^k$$

and a sequence of critical values

$$\lambda_{i+1} = \min\{\lambda \mid K_{\mathcal{X}^i} + \lambda \mathcal{H}^i \text{ is nef over } \mathbb{C}\}$$

with $q + 1 = \lambda_0 \ge \lambda_1 \ge \cdots \ge \lambda_k > \lambda_{k+1} = 1$. For any $\lambda_i \ge \lambda \ge \lambda_{i+1}$, we let $\mathcal{H}^i$ be the pushforward of $\mathcal{H}$ to $\mathcal{X}^i$ and set

$$(128) \qquad \mathcal{L}^i_\lambda = \frac{1}{\lambda - 1}(K_{\mathcal{X}^i} + \lambda \mathcal{H}^i) = \frac{1}{\lambda - 1}(K_{\mathcal{X}^i} + \mathcal{H}^i) + \mathcal{H}^i =: \frac{1}{\lambda - 1}E + \mathcal{H}^i.$$

Write $E = \sum_{j=1}^k e_j \mathcal{X}_{0,j}^i$ with $e_1 \le e_2 \le \cdots \le e_k$. Then we have $(d/d\lambda)\mathcal{L}^i_\lambda = -(1/(\lambda-1)^2)E$ and

$$\boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}^i, \mathcal{L}^i, a\eta^i/d_1) = \frac{a\lambda}{\lambda - 1} e_1.$$

So we can again use (124) to calculate

$$\frac{d}{d\lambda} \boldsymbol{H}^{\mathrm{NA}}\left(\mathcal{X}^i, \mathcal{L}^i_\lambda, \frac{a\eta^i}{d_1}\right) = -\frac{a}{d_1(\lambda-1)^2} e_1 + \frac{a}{d_1(\lambda-1)^2} \frac{\sum_i e_i \boldsymbol{Q}_i^{(ad_1^{-1})}}{\boldsymbol{Q}^{(ad_1^{-1})}}$$

$$= \frac{a}{d_1(\lambda-1)^2} \frac{\sum_i (e_i - e_1) \boldsymbol{Q}_i^{(a)}}{\boldsymbol{Q}^{(a)}} \ge 0.$$

The last identity holds only if $e_i \equiv e_1$, which implies $(\mathcal{X}^{\mathrm{lc}}, \mathcal{L}^{\mathrm{lc}}) \cong (\mathcal{X}^{\mathrm{ac}}, \mathcal{L}^{\mathrm{ac}} + e_1 \mathcal{X}_0^{\mathrm{ac}})$.

Step 4 With the test configuration $(\mathcal{X}^{\mathrm{ac}}, \mathcal{L}^{\mathrm{ac}})$ obtained from Step 2, there exists a $d_2 \in \mathbb{Z}_{>0}$ and a projective birational $T_{\mathbb{C}} \times \mathbb{C}^*$–equivariant birational map $(\mathcal{X}^{\mathrm{ac}})^{(d_2)} \dashrightarrow \mathcal{X}^s$ over $\mathbb{P}^1$ such that $(\mathcal{X}^s, -K_{\mathcal{X}^s})$ is a special test configuration and

$$(129) \qquad \boldsymbol{H}^{\mathrm{NA}}\left(\mathcal{X}^{\mathrm{ac}}, \mathcal{L}^{\mathrm{ac}}, \frac{a\eta}{d_1 d_2}\right) \ge \boldsymbol{H}^{\mathrm{NA}}\left(\mathcal{X}^s, \mathcal{L}^s, \frac{a\eta^s}{d_1 d_2}\right).$$

As in [56], this is achieved by doing a base change and running an MMP. Let $E = -K_{\mathcal{X}^s/\mathbb{P}^1} - (-K_{\mathcal{X}'/\mathbb{P}^1})$. Then $E \geq 0$ by the negativity lemma. So $\mathcal{L}'_\lambda = -K_{\mathcal{X}'/\mathbb{P}^1} + \lambda E$, and

$$(130) \qquad \mathrm{lct}\left(\mathcal{X}', \mathcal{L}'_\lambda, \frac{a\eta'}{d_1 d_2}\right) = \frac{a}{d_1 d_2}\lambda e_1.$$

So, as before, we get

$$\frac{d}{d\lambda} \boldsymbol{H}^{\mathrm{NA}}\left(\mathcal{X}', \mathcal{L}'_\lambda, \frac{a\eta'}{d_1 d_2}\right) = \frac{a}{d_1 d_2} e_1 - \frac{a}{d_1 d_2} \frac{\sum_i e_i \boldsymbol{Q}_i^{(ad_1^{-1})}}{\boldsymbol{Q}^{(ad_1^{-1})}} = -\frac{a}{d_1 d_2} \frac{\sum_i (e_i - e_1) \boldsymbol{Q}_i^{(ad_1^{-1})}}{\boldsymbol{Q}^{(ad_1^{-1})}} \leq 0.$$

The last identity holds only if $e_i \equiv e_1$ which implies $(\mathcal{X}^{\mathrm{ac}}, \mathcal{L}^{\mathrm{ac}}) \cong (\mathcal{X}^s, \mathcal{L}^s)$. $\qquad \square$

Corollary 3.6 *We have the identity*

$$(131) \qquad h(X) = \inf_{(\mathcal{X}, \mathcal{L}, a\eta) \text{ special}} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta).$$

Lemma 3.7 *For any normal test configuration* $(\mathcal{X}, \mathcal{L}, \eta)$, *there exists a unique* $a_* > 0$ *such that*

$$(132) \qquad \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a_*\eta) = \inf_{c > 0} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) =: \boldsymbol{H}_*^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}).$$

As a consequence, we have

$$(133) \qquad h(X) = \inf_{(\mathcal{X}, -K_{\mathcal{X}}) \text{ special}} \boldsymbol{H}_*^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}).$$

Proof By taking normalization of a fiber product, without loss of generality we can assume that $\mathcal{X}$ dominates $X_{\mathbb{C}} := X \times \mathbb{C}$ by a $\mathbb{C}^*$–equivariant birational morphism $\rho \colon \mathcal{X} \to X_{\mathbb{C}}$.

Choose a $\mathbb{C}^*$–equivariant resolution of singularities $\mu \colon \widetilde{\mathcal{X}} \to \mathcal{X}$ such that the pair $(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}_0^{\mathrm{red}})$ has simple normal crossing singularities. Set $\widetilde{\rho} = \rho \circ \mu$. Then we can write

$$K_{\widetilde{\mathcal{X}}} = \widetilde{\rho}^* K_{X_{\mathbb{C}}} + \sum_i a_i E_i + \sum_j a'_j E'_j, \quad \pi^* \mathcal{X}_0 = \sum_i b_i E_i, \quad \mu^* \mathcal{L} = \widetilde{\rho}^*(-K_{X_{\mathbb{C}}}) + \sum_i c_i E_i,$$

where $\{E_i\}$ are irreducible components of $\widetilde{\mathcal{X}}_0$ and $\{E'_j\}$ are horizontal exceptional divisors. Then we have the identity (see [19, Proposition 7.29])

$$\boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}) = \mathrm{lct}(\mathcal{X}, -(K_{\mathcal{X}} + \mathcal{L}); \mathcal{X}_0) - 1$$
$$= \min_i \left(b_i^{-1} A_{(X \times \mathbb{C}, X \times \{0\})}(E_i) + b_i^{-1} \mathrm{ord}_{E_i}\left(\sum_k c_k E_k\right) \right) = \min_i \frac{1 + a_i + c_i}{b_i} - 1.$$

Because $\boldsymbol{H}^{\mathrm{NA}}$ is translation-invariant, by adding a multiple of $\mathcal{X}_0$ to $\mathcal{L}$ we can normalize $\phi = \phi_{\mathcal{F}}$ to satisfy $\boldsymbol{L}^{\mathrm{NA}}(\phi) = 0$. So we get

$$(134) \qquad c_i \geq b_i - 1 - a_i$$

and, without loss of generality, $c_1 = b_1 - 1 - a_1$. So

$$(135) \qquad \lambda_{\min} = \min_i \frac{c_i}{b_i} \leq \frac{c_1}{b_1} = 1 - \frac{a_1 + 1}{b_1} = 1 - A_{X_{\mathbb{C}}}(b_1^{-1} \mathrm{ord}_{E_1}) = -A_X(v_{E_1}) \leq 0,$$

where $v_{E_1} := r(b_1^{-1}\mathrm{ord}_{E_1})$ is the restriction of the valuation ord_{E_1} to the function field $\mathbb{C}(X)$. Here we used the identity between log discrepancies from [19, Proposition 4.11] and the assumption that X has log terminal singularities.

Set $\mathcal{F} = \mathcal{F}_{(\mathcal{X},\mathcal{L},\eta)}$. Then according to (43), we have $\mathcal{F}_{(\mathcal{X},\mathcal{L},a\eta)} = a\mathcal{F}_{(\mathcal{X},\mathcal{L},\eta)}$. Moreover, by (74), (76) and (90), we get the expression

$$f(a) := \boldsymbol{H}^{\mathrm{NA}}(a\mathcal{F}) = a\boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}) + \log\left(\frac{1}{V}\int_{\lambda_{\min}}^{+\infty} e^{-a\lambda}\,\mathrm{DH}(\mathcal{F})\right).$$

By Lemma 2.36, $f(a)$ is convex in $a \in [0, +\infty)$. If $\lambda_{\min} < 0$, then $f(a)$ diverges to $+\infty$ as $a \to +\infty$ by the above expression. So $f(a)$ admits a unique minimum over $[0, +\infty)$.

If $\lambda_{\min} = 0$, then by (135) $A_X(v_{E_1}) = 0$, which implies that v_{E_1} is trivial so that

$$\lambda_{\max} = \mathrm{ord}_{E_1}\left(\sum_k c_k E_k\right) = c_1 = 0 = \lambda_{\min}.$$

See [19, Theorem 5.16]. This implies that the normal test configuration $(\mathcal{X}, \mathcal{L})$ is equivalent to a trivial test configuration and hence $\boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) \equiv 0$. $\square$

4 A minimization problem for real valuations

In this section, we will introduce a minimization problem for valuations analogous to the normalized volume functional in the local setting [51].

Definition 4.1 For any $v \in \mathrm{Val}(X)$, define

$$(136) \qquad \widetilde{\beta}(v) = \begin{cases} A_X(v) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}_v) & \text{if } A_X(v) < +\infty, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that by integration by parts we have

$$(137) \qquad e^{-\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}_v)} = \frac{1}{V}\int_{\mathbb{R}} e^{-x}\,\mathrm{DH}(\mathcal{F}_v) = \frac{1}{V}\int_0^{+\infty} e^{-x}(-d\,\mathrm{vol}(\mathcal{F}_v^{(x)}))$$

$$= 1 - \frac{1}{V}\int_0^{+\infty} \mathrm{vol}(\mathcal{F}_v^{(x)}R_\bullet)e^{-x}\,dx \le 1,$$

with identity if and only if v is trivial. So we can rewrite $\widetilde{\beta}(v)$ as

$$(138) \qquad \widetilde{\beta}(v) = A_X(v) + \log\left(1 - \frac{1}{V}\int_0^{+\infty} e^{-x}\,\mathrm{vol}(\mathcal{F}_v^{(x)}R_\bullet)\,dx\right).$$

Lemma 4.2 *For any $v \in X_{\mathbb{Q}}^{\mathrm{div}}$, we have the inequality*

$$(139) \qquad \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}_v) \le \widetilde{\beta}(v).$$

Moreover, if $(\mathcal{X}, -K_{\mathcal{X}}, a\eta)$ is a special test configuration, then equality holds for $v = av_{x_0} = a \cdot r\,(\mathrm{ord}_{x_0})$. (See Definition 2.12.)

Proof The inequality follows immediately from

(140)
$$\inf_w (A(w) + \phi_v(w)) \leq A(v) + \phi_v(v) = A(v).$$

When $(\mathcal{X}, -K_{\mathcal{X}}, a\eta)$ is a special test configuration and $v = a v_{\mathcal{X}_0}$, then

(141)
$$\boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}, -K_{\mathcal{X}}, a\eta) = a\boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}, -K_{\mathcal{X}}, \eta) = a(\mathrm{lct}(\mathcal{X}; \mathcal{X}_0) - 1) = 0.$$

On the other hand, by (46),

(142)
$$\boldsymbol{L}^{\mathrm{NA}}(\mathcal{X}, -K_{\mathcal{X}}, a\eta) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}_v(-A(v))) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}_v) - A(v).$$

So we get

(143)
$$\boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}_v) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}_v) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}_v) = A(v) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(v) = \widetilde{\beta}(v). \qquad \square$$

Lemma 4.3 *For any $\phi = \phi_{\mathcal{F}}$ and $v \in X_{\mathbb{Q}}^{\mathrm{div}}$, we have the inequality*

(144)
$$\widetilde{\boldsymbol{S}}(v) + \phi(v) \geq \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\phi).$$

Proof We use the same argument as in [52, Section 4.1]. Set $\gamma = \phi(v)$. Then by the argument there, we have $\lambda_{\min} = \lambda_{\min}(\mathcal{F}) \leq \gamma$ and we can then estimate

$$
\begin{aligned}
e^{-\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\phi)} = \boldsymbol{Q}(\phi) &= \frac{1}{V}\int_{\mathbb{R}} e^{-x}(-d\mathrm{vol}(\mathcal{F}^{(x)})) = e^{-\lambda_{\min}} - \frac{1}{V}\int_{\lambda_{\min}} e^{-x}\,\mathrm{vol}(\mathcal{F}^{(x)} R_\bullet)\,dx \\
&\geq e^{-\gamma} - \frac{1}{V}\int_{\gamma}^{+\infty} e^{-x}\,\mathrm{vol}(\mathcal{F}^{(x)} R_\bullet)\,dx \geq e^{-\gamma} - \frac{1}{V}\int_{\gamma}^{+\infty} e^{-x}\,\mathrm{vol}(\mathcal{F}_v^{(x-a)})\,dx \\
&= e^{-\gamma} - e^{-\gamma}\frac{1}{V}\int_0^{+\infty} e^{-x}\,\mathrm{vol}(\mathcal{F}_v^{(t)})\,dt = e^{-\gamma}\frac{1}{V}\int_0^{+\infty} e^{-x}(-d\mathrm{vol}(\mathcal{F}_v^{(t)})) \\
&= e^{-\phi(v)}e^{-\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(v)}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square
\end{aligned}
$$

Proposition 4.4 *For any $\mathbb{Q}$–Fano variety, we have the identity*

(145)
$$h(X) = \inf_{v \in X_{\mathbb{Q}}^{\mathrm{div}}} \widetilde{\beta}(v).$$

Proof For any test configuration $(\mathcal{X}, \mathcal{L}, a\eta)$, by Theorem 3.5 there exists a special test configuration $(\mathcal{X}^s, \mathcal{L}^s, a^s \eta^s)$ such that

(146)
$$\boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}, \mathcal{L}, a\eta) \geq \boldsymbol{H}^{\mathrm{NA}}(\mathcal{X}^s, \mathcal{L}^s, a^s \eta^s) = \widetilde{\beta}(a^s v_{\mathcal{X}_0^s}).$$

The last identity is from Lemma 4.2. This together with Corollary 3.6 implies identity (145).

Alternatively, recall that $\boldsymbol{L}^{\mathrm{NA}}(\phi) = \inf_{v \in X_{\mathbb{Q}}^{\mathrm{div}}}(A_X(v) + \phi(v))$. So for any $\epsilon > 0$ we can choose v such that $A_X(v) + \phi(v) < \boldsymbol{L}^{\mathrm{NA}}(\phi) + \epsilon$. We can then use v in (144) to get

$$\boldsymbol{L}^{\mathrm{NA}}(\phi) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\phi) \geq A_X(v) + \phi(v) - \epsilon - (\phi(v) + \widetilde{\boldsymbol{S}}(v)) = \widetilde{\beta}(v) - \epsilon.$$

Since ϵ is arbitrary, we can use (97) to get the identity (145). $\qquad\qquad\qquad\qquad \square$

With the identity (145) and Proposition 2.32, we get:

Corollary 4.5 *For any* $\mathbb{Q}$*–Fano variety, we have the equality*

$$(147) \qquad\qquad h(X) = \inf_{\mathcal{F}} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}).$$

The next result should be compared to Lemma 3.7.

Proposition 4.6 *For any* $v \in \overset{\circ}{\mathrm{Val}}(X)$*, there exists a unique* $a_* = a_*(v) \geq 0$ *such that*

$$(148) \qquad\qquad \widetilde{\beta}(a_* v) = \inf_{a>0} \widetilde{\beta}(av) =: \widetilde{\beta}_*(v).$$

When $\beta(v) \geq 0$*, then* $a_* = 0$*, so that* $a_* v$ *is the trivial valuation and* $\widetilde{\beta}_*(v) = 0$*. Otherwise,* $a_*(v) > 0$ *and* $\widetilde{\beta}_*(v) < 0$.

Proof Consider the function defined on $\mathbb{R}_{\geq 0}$ by

$$(149) \qquad f(a) = A(av) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(av) = aA(v) + \log\left(\frac{1}{V} \int_0^{+\infty} e^{-x} \, \mathrm{DH}(\mathcal{F}_{av}) \right)$$

$$= aA(v) + \log\left(\frac{1}{V} \int_0^{+\infty} e^{-ax} \, \mathrm{DH}(\mathcal{F}_v) \right).$$

We will show that $a \mapsto f(a)$ is convex and goes to $+\infty$ as $a \to +\infty$. Now

$$f'(a) = A(v) - \frac{\int_0^{+\infty} x e^{-ax} \, \mathrm{DH}(\mathcal{F}_v)}{\int_0^{+\infty} e^{-ax} \, \mathrm{DH}(\mathcal{F}_v)},$$

$$f''(a) = \frac{\int x^2 e^{-ax} \, \mathrm{DH}}{\int e^{-ax} \, \mathrm{DH}} - \frac{\left(\int x e^{-ax} \, \mathrm{DH} \right)^2}{\left(\int e^{-ax} \, \mathrm{DH} \right)^2} = \| x - \bar{x} \|_{L^2(dv)}^2 \geq 0,$$

where

$$(150) \qquad\qquad dv = \frac{e^{-ax} \, \mathrm{DH}}{\int e^{-ax} \, \mathrm{DH}} \quad \text{and} \quad \bar{x} = \int x \, dv.$$

So $f''(a) = 0$ if and only if av is trivial. Moreover, $f'(0) = A(v) - (1/V) \int_0^{+\infty} x \, \mathrm{DH}(\mathcal{F}_v) = \beta(v)$.

On the other hand, $f(0) = 0$ and we claim that $\lim_{a \to +\infty} f(a) = +\infty$, which then implies the statement. To prove this divergence, we set $g(x) = V^{-1/n} \mathrm{vol}(\mathcal{F}^{(x)} R_\bullet)^{1/n}$. Then $g(x)$ is decreasing, and concave on $[0, \lambda_{\max}]$ by Theorem 2.5. As a consequence, the subset $\{x \in \mathbb{R}_{\geq 0} \mid g'(x) \text{ exists}\}$ is dense in $\mathbb{R}_{\geq 0}$, by Aleksandrov's differentiability theorem for concave functions. Fix $0 < \epsilon \ll \lambda_{\max}$ such that $g'(\epsilon)$ exists and $g(\epsilon) < g(0) = 1$. Setting $C = -g'(\epsilon) > 0$ and $T = (1 + C\epsilon)/C$, define a function

$$(151) \qquad\qquad \hat{g}(x) = \begin{cases} 1 & \text{if } x \in [0, \epsilon], \\ 1 + C\epsilon - Cx & \text{if } x \in (\epsilon, T], \\ 0 & \text{if } x \in (T, +\infty). \end{cases}$$

Then $\hat{g}(x) \geq g(x)$ over $[0, +\infty)$, by concavity. Then we calculate to get

$$(152) \qquad\qquad a \int_0^{+\infty} \hat{g}^n(x) e^{-ax} \, dx = 1 - nC m_{n-1},$$

where $m_k = \int_\epsilon^T (1 + C\epsilon - Cx)^k e^{-ax}\, dx$ satisfies

$$m_k = \frac{1}{a}e^{-a\epsilon} - \frac{kC}{a}m_{k-1} = \frac{1}{a}e^{-a\epsilon} - \frac{kC}{a}\left(\frac{1}{a}e^{-a\epsilon} - \frac{(k-1)C}{a}m_{k-2}\right).$$

Using induction we get $m_{n-1} = a^{-1}e^{-a\epsilon}(1 + O(a^{-1}))$. So

$$e^{-\widetilde{\boldsymbol{S}}^{\mathrm{NA}}}(\mathcal{F}_{av}) = 1 - a\int_0^{+\infty} g^n(x)e^{-ax}\,dx \geq 1 - a\int_0^{+\infty} \hat{g}^n(x)e^{-ax}\,dx$$

$$= nCm_{n-1} = nCa^{-1}e^{-a\epsilon}(1 + O(a^{-1})).$$

So we get $-\widetilde{\boldsymbol{S}}^{\mathrm{NA}}(\mathcal{F}_{av}) \geq -\log a - a\epsilon + O(1)$, giving

(153) $$f(a) = \widetilde{\beta}(av) \geq (A(v) - \epsilon)a - \log a + O(1),$$

which approaches $+\infty$ as $a \to +\infty$ if we choose $0 < \epsilon < A(v)$. $\square$

Remark 4.7 By the above proof, we get an estimate: for any $C_1 > 0$, there exists a $C_2 = C_2(C_1, v) > 0$ such that for any $w \in \mathrm{Val}(X)$ with $w \leq C_1 v$, we have

(154) $$a_*(w) \leq \frac{C_2}{A(v)}.$$

Corollary 4.8 *We always have $h(X) \leq 0$, with equality holding if and only if $h(X) = 0$.*

Proof By [37; 50], X is K–semistable if and only if $\beta(v) \geq 0$, which implies $\widetilde{\beta}_*(v) = 0$. If X is not K–semistable then there exists v' such that $\beta(v') < 0$. By Proposition 4.6, we then have $\widetilde{\beta}_*(v') < 0$, which implies $h(X) < 0$. $\square$

Lemma 4.9 *If v computes $h(X)$, then v is the unique valuation, up to rescaling, that computes $\mathrm{lct}(\mathfrak{a}_\bullet(v))$.*

Proof Recall that $\mathrm{lct}(\mathfrak{a}_\bullet) = \inf_w A(w)/w(\mathfrak{a}_\bullet(w))$. For any $w \in \mathrm{Val}(X)$, assume that $w(\mathfrak{a}_\bullet(v)) = a > 0$. Then $a^{-1}w \geq v$. By Proposition A.1, the function

$$w \mapsto \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(w) = -\log\frac{1}{V}\int_{\mathbb{R}} e^{-\lambda}\,\mathrm{DH}(\mathcal{F}_w)$$

is strictly increasing on $\mathrm{Val}(X)$. So we use the assumption to get

$$\frac{A(w)}{w(\mathfrak{a}_\bullet(v))} = A(a^{-1}w) = A(a^{-1}w) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(a^{-1}w) + \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(a^{-1}w)$$

$$\geq A(v) - \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(v) + \widetilde{\boldsymbol{S}}^{\mathrm{NA}}(v) = A(v)$$

$$= \frac{A(v)}{v(\mathfrak{a}_\bullet(v))}.$$

When the equality holds, then $a^{-1}w = v$. $\square$

We now observe that the method developed in [14] can be used to prove:

Theorem 4.10 *For any $\mathbb{Q}$–Fano variety, there exists a minimizing valuation of $\widetilde{\beta}$ which is quasi-monomial.*

Since the argument is almost verbatim to [14] except for the continuity property of $\widetilde{\beta}$, we just give a sketch of key points and explain the required continuity of $\widetilde{\beta}$ in Section 8. Without the properties of $\widetilde{\beta}(S)$ explained in Section 8, the existence of a valuation calculating $h(X)$ (but without the quasimonomial property) can also be obtained using the argument in [11, Section 6].

Proof By Corollary 3.6, $h(X) = \inf_E \widetilde{\beta}_*(E)$, where E ranges over prime divisors over X that induce special test configurations of $(X, -K_X)$. By [14, Theorem A.2], we know that such an E is an lc place of an N–complement D of X, where N depends only on the dimension n (this depends on the deep result of Birkar about the boundedness of $\mathbb{Q}$–complements). So we have

$$(155) \qquad\qquad h(X) = \inf_v \widetilde{\beta}_*(v),$$

where v ranges over all divisorial valuations that are lc places of an N–complement. For such a valuation v, there exists a $D \in (1/N)|-NK_X|$ such that (X, D) is lc and $A_{(X,D)}(v) = 0$. We then parametrize such $\mathbb{Q}$–divisors as in [14, Proof of Theorem 4.5]. Set $W = \mathbb{P}(H^0(X, \mathcal{O}_X(-NK_X)))$ and denote by H the universal divisor on $X \times W$ parametrizing divisors in $|-NK_X|$ and set $D := (1/N)H$. By the lower semicontinuity of log canonical thresholds, the locus $Z = \{w \in W \mid \mathrm{lct}(X_w; D_w) = 1\}$ is locally closed in W. For each $z \in Z$, set $b_z := \inf_v \widetilde{\beta}(v)$, where v ranges over all $v \in \mathrm{Val}(X)$ with $A_{(X,D_z)}(v) = 0$.

Let $g \colon Y_z \to X$ be a log resolution of (X, D_z). Write $K_Y + D_{Y_z} = g^*(K_X + D_z)$. Consider the section of the simplicial cone, $S := \mathrm{QM}(Y_z, D_{Y_z}) \bigcap \{v \in \mathrm{Val}(X) \mid A(v) = 1\}$. By Proposition 4.6, we know that for each $v \in S$ there exists $a_*(v)$ such that $\inf_{a>0} \widetilde{\beta}(av) = \widetilde{\beta}(a_*(v)v) =: \widetilde{\beta}_*(v)$. By Izumi's estimate (see [48, Example 11.3.9; 44, Proposition 5.10] for the smooth case, and [51, Section 3] in the klt case), we know that there exists $C_1 > 0$ such that for any $v \in S$ we have $v \le C_1 \cdot \mathrm{ord}_F$, where $F = \bigcap_i D_{Y_z,i}$. Now by the proof of Proposition 4.6 (see Remark 4.7), we know that $a_*(v)$ is uniformly bounded for any $v \in S$. By Proposition A.2, we know that $v \mapsto \widetilde{\beta}(v)$ is continuous on $\mathrm{QM}(Y_z, D_{Y_z})$ and hence is uniformly continuous over compact subsets. We then get the continuity of $v \mapsto \widetilde{\beta}_*(v)$ over the compact set S. So we know that there exists $v_z^* \in S$ such that $\widetilde{\beta}_*(v_z^*) = \min_{v \in S} \widetilde{\beta}_*(v)$ and $a_*(v_z^*) \cdot v_z^*$ is then a minimizer of $\widetilde{\beta}$ over $\mathrm{QM}(Y_z, D_{Y_z})$.

Then as [14, Proof of Theorem 4.5], choose a locally closed decomposition $Z = \bigcup_{i=1}^r Z_i$ so that Z_i is smooth and there is an étale map $Z_i' \to Z_i$ such that $(X_{Z_i'}, D_{Z_i'})$ admits fiberwise log resolutions. By the same arguments as [14, Proof of Propositions 4.1 and 4.2], which depend on the deformation invariance of log plurigenera in the work of Hacon, McKernan and Xu, we know that b_z is independent of $z \in Z_i$. So b_z takes finitely many values and there is a $z_0 \in Z$ such that $h(X) = \min_{z \in Z} b_z = b_{z_0}$ is computed by $v_{z_0}^*$. □

As in the case of normalized volume, we expect the following:

Conjecture 4.11 *The minimizer v_* is unique, and is special, which means that $\mathcal{F}_{v_*}$ is a special $\mathbb{R}$–test configuration.*

Remark 4.12 As [11, Proposition 4.11], using Lemma 4.9 one can show that any divisorial (ie rational rank one) minimizing valuation is primitive and plt.

Besides the case of stability threshold treated in [14], in the local setting of normalized volumes, the existence of quasimonomial minimizers is also known thanks to the work of Blum [10] and Xu [78]. Moreover, one might also be able to adapt the techniques in Xu and Zhuang [80] to the current global setting to prove the uniqueness of minimizing valuations.[3] We will prove in Section 6 the uniqueness of special minimizers (in a similar spirit to the work in [58; 57; 55]).

5   Initial term degeneration of filtrations

Let $\mathcal{F}_0$ be a special $\mathbb{R}$–test configuration of $(X, -K_X)$ with central fiber $(W := \mathrm{Proj}(\mathrm{Gr}(\mathcal{F}_0)), \xi_0 := \xi_{\mathcal{F}_0})$. Let $\mathcal{F}_1$ be another filtration of R. We define a filtration on

$$(156) \qquad R' := R(W, -K_W) = \bigoplus_{m \geq 0} \bigoplus_{\lambda \in \Gamma(\mathcal{F}_0)} t^{-\lambda} \mathcal{F}_0^\lambda R_m / \mathcal{F}_0^{>\lambda} R_m =: \bigoplus_{m \geq 0} R'_m$$

in the following way. Recall that we can write

$$(157) \qquad R'_m = \bigoplus_{\alpha \in M_\mathbb{Z}} t^{-\langle \alpha, \xi_0 \rangle} \mathcal{F}_0^{\langle \alpha, \xi_0 \rangle} R_m / \mathcal{F}_0^{>\langle \alpha, \xi_0 \rangle} R_m.$$

For any $f \in R_m$, set

$$(158) \qquad \mathbf{in}_{\mathcal{F}_0}(f) = (t^{-\langle \alpha, \xi \rangle} \bar{f})(0) =: f' \in \mathcal{F}_0^{\langle \alpha, \xi_0 \rangle} R_m / \mathcal{F}_0^{>\langle \alpha, \xi_0 \rangle} R_m, \quad \text{where } \langle \alpha, \xi_0 \rangle = v_{\mathcal{F}_0}(f).$$

For any $\lambda \in \mathbb{R}$, take the Gröbner base-type degeneration

$$(159) \qquad \mathcal{F}_1'^\lambda R'_m = \mathrm{Span}_\mathbb{C}\{\mathbf{in}_{\mathcal{F}_0}(f) \mid f \in \mathcal{F}_1^\lambda R_m\} \subseteq R'_m.$$

Note that because R' is integral, $\mathbf{in}_{\mathcal{F}_0}(fg) = \mathbf{in}_{\mathcal{F}_0}(f) \cdot \mathbf{in}_{\mathcal{F}_0}(g)$ if $f \in R_{m_1}$ and $g \in R_{m_2}$. So in this way, we get a $\mathbb{T}_0$–equivariant filtration

$$(160) \qquad \mathcal{F}_1'^\lambda R'_m = \bigoplus_{\alpha \in \mathbb{Z}^{r_0}} \mathcal{F}_1'^\lambda R'_{m,\alpha}.$$

There is an equivalent way to describe $\mathcal{F}_1'^\lambda R'_m$, as follows. For any $f' \in R'_{m,\alpha}$, we choose $f \in R_m$ such that $f' = t^{-\langle \alpha, \xi_0 \rangle} \bar{f}(0)$. Then we have

$$(161) \qquad \mathcal{F}_1'^\lambda R'_{m,\alpha} = \{f' \in R'_{m,\alpha} \mid f + h \in \mathcal{F}_1^\lambda R_m \text{ for some } h \in \mathcal{F}_0^{>\langle \alpha, \xi_0 \rangle} R_m\}.$$

This is well defined since f is determined up to addition by elements from $\mathcal{F}_0^{>\langle \alpha, \xi_0 \rangle} R_m$.

Note that this construction allows us to find a basis $\mathcal{B} = \{s_1, \dots, s_{N_m}\}$ of R_m that is compatible with both $\mathcal{F}_0 R_m$ and $\mathcal{F}_1 R_m$. Recall that this means that for any $\lambda \in \mathbb{R}$ and $i = 0, 1$, there exists a subset of $\mathcal{B}$ which depends on λ and i and is a basis of $\mathcal{F}_i^\lambda R_m$. To find such a basis, we can first find a basis $\mathcal{B}'_\alpha$ of $R'_{m,\alpha}$ which is compatible with $\mathcal{F}_1' R_{m,\alpha}$. Then $\mathcal{B} = \bigcup_\alpha \mathcal{B}_\alpha =: \{f'_1, \dots, f'_{N_m}\}$ is a basis compatible with

---

[3]This has indeed been recently achieved in [15].

both $\mathcal{F}'_1 R_m$ and $\mathcal{F}'_{\mathrm{wt}_{\xi_0}} R'_m$. For each $f'_k \in R'_{m,\alpha_k}$, there exists $\lambda_k \in \mathbb{R}$ such that $f'_k \in \mathcal{F}'^{\lambda_k}_1 R'_m \setminus \mathcal{F}'^{>\lambda_k}_1 R'_m$. Then by (161), there exists $h_k \in \mathcal{F}^{>\langle \alpha_k, \xi \rangle}_0 R_m$ such that $s_k := f_k + h_k \in \mathcal{F}^{\lambda_k}_1 R_m$. Moreover, we have $s_k \notin \mathcal{F}^{>\lambda_k}_1 R_m$ since otherwise $\mathbf{in}(s_k) = \mathbf{in}(f_k) = f'_k \in \mathcal{F}'^{>\lambda_k} R'_m$. It is easy to verify that $\{s_k\}$ is the desired basis. So the relative successive minima of $\mathcal{F}_1$ with respect to $\mathcal{F}_0$ (see [20]) is given by the set $\{\lambda_k - \langle \alpha_k, \xi_0 \rangle\}$, which is the same as the relative successive minima of $\mathcal{F}'_1 := \mathcal{F}'_1 R'$ with respect to $\mathcal{F}'_0 := \mathcal{F}'_{\mathrm{wt}_{\xi_0}} R'$. This immediately proves a useful fact:

Lemma 5.1 *With the above constructions and notation, we have the identity*

$$(162) \qquad\qquad d^X_2(\mathcal{F}_0, \mathcal{F}_1) = d^W_2(\mathcal{F}'_0, \mathcal{F}'_1).$$

Since the initial term degeneration does not change the dimension of vector spaces, it is clear that the successive minima of $\mathcal{F}_1$ and $\mathcal{F}'_1$ coincide. As a consequence, we get

$$(163) \qquad\qquad \widetilde{\boldsymbol{S}}^{\mathrm{NA}}_X(\mathcal{F}_1) = \widetilde{\boldsymbol{S}}^{\mathrm{NA}}_W(\mathcal{F}'_1).$$

On the other hand, consider the $\mathbb{T}_0$–equivariant graded filtration of the Rees algebra $\mathcal{R}' := \mathcal{R}(\mathcal{F}_0)$ (see (30)) given by

$$(164) \qquad\qquad \mathcal{F}'^\lambda \mathcal{R}'_{m,\alpha} = \{ s = t^{-\langle \alpha, \xi \rangle} \overline{f} \in \mathcal{R}'_{m,\alpha} \mid t^{-\lambda} \overline{f} \in \mathcal{R}(\mathcal{F}_1) \}.$$

Then $\mathcal{F}'\mathcal{R}'$ coincides with $\mathcal{F}R$ on the generic fiber and coincides with $\mathcal{F}'R'$ on the central fiber. By the lower semicontinuity of lct for a family, it is easy to see that $\hat{\boldsymbol{L}}^{\mathrm{NA}}$ in (80) is also lower semicontinuous for a family. This is standard if $\mathcal{F}_0$ has rank one, which corresponds to a special test configuration; see [47, Lemma 8.1; 13, Proof of Lemma 6.5]. In general, one can restrict to a generic curve passing through 0 in the family in Teissier's construction in the paragraph above Lemma 2.11; alternatively, see Remark 6.2. So we can get

$$(165) \qquad \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}_1) = \hat{\boldsymbol{L}}^{\mathrm{NA}}_X(\mathcal{F}_1) \geq \hat{\boldsymbol{L}}^{\mathrm{NA}}_W(\mathcal{F}'_1) = \boldsymbol{L}^{\mathrm{NA}}(\mathcal{F}'_1),$$

where the first and the last identity come from Proposition 2.32. Combining the above discussion, we get the inequality

$$(166) \qquad\qquad \boldsymbol{H}^{\mathrm{NA}}_X(\mathcal{F}_1) \geq \boldsymbol{H}^{\mathrm{NA}}_W(\mathcal{F}'_1).$$

Theorem 5.2 *Assume v induces a special $\mathbb{R}$–test configuration $\mathcal{F}_v$ of X. Then v is a minimizer of $\widetilde{\beta}$ over $\mathrm{Val}(X)$ if and only if v is Ding-semistable (or equivalently K–semistable).*

Proof For simplicity of notation, set $\mathcal{F}_0 = \mathcal{F}_v$ and $(W, \xi_0) := (X_{\mathcal{F}_v,0}, \xi_{\mathcal{F}_0})$ and let $\mathbb{T}_0$ be the torus generated by ξ_0.

We first prove that minimizer is Ding-semistable. Suppose (W, ξ_0) is not Ding-semistable. Then by Theorem 2.46 from [39], there exists a $\mathbb{T}$–equivariant special test configuration $(\mathcal{W}, -K_{\mathcal{W}})$ of $(W, -K_W)$ with central fiber $Y := \mathcal{W}_0$ such that

$$(167) \qquad\qquad \boldsymbol{D}^{\mathrm{NA}}_g(\mathcal{W}, -K_{\mathcal{W}}) = \mathrm{Fut}_{Y,\xi}(\eta) < 0.$$

We can now construct a family of valuations $\{v_\epsilon\}$ such that v_ϵ induces special test configurations with central fiber Y and corresponds to a vector field $\xi_\epsilon = \xi_0 + \epsilon\eta$ on Y. This can be done by using the cone construction to reduce to the situation in [58, Section 6] or [57, Proof of Theorem 2.64]. Alternatively, one can use an argument involving the Hilbert scheme as in [55, Proof of Lemma 3.1].

Here we will use the Chow variety to explain this construction. Recall that the Chow point of a cycle $Z \subset \mathbb{P}^{N-1}$ of degree d and dimension n corresponds to a divisor in the Grassmannian $\mathrm{Gr}(N-n-1, \mathbb{C}^N)$ which is the zero scheme of a section:

$$\mathrm{CH}(Z) \in H^0(\mathrm{Gr}(N-n-1, \mathbb{C}^N), \mathcal{O}(d)) =: \mathbb{M}.$$

$\mathrm{CH}(Z)$ is determined up to rescaling and we call it the Chow coordinate of Z. Let $\mathrm{CH}(X)$, $\mathrm{CH}(W)$ and $\mathrm{CH}(Y)$ be the Chow coordinates of X, W and Y, respectively. Denote by $[\mathrm{CH}(X)]$ the Chow point of X in the projectivization $\mathbb{P}(\mathbb{M})$, and similarly for Y and W. Since the $\mathbb{T}$–action on $\mathbb{P}^{N-1}$ induces a weight decomposition $\mathbb{M} = \bigoplus_\alpha \mathbb{M}_\alpha$, we have

$$(168) \qquad \lim_{s\to+\infty} \sigma_\xi(s) \circ [\mathrm{CH}(X)] = [\mathrm{CH}(W)] \quad \text{and} \quad \lim_{s\to+\infty} \sigma_\eta(s) \circ [\mathrm{CH}(W)] = [\mathrm{CH}(Y)].$$

If we set

$$(169) \qquad \mathrm{CW}_\xi(X) = \min\{\langle \alpha, \xi \rangle \mid \mathrm{CH}(X)_\alpha \neq 0\} \quad \text{and} \quad \mathrm{CW}_\eta(W) = \min\{\langle \alpha, \xi \rangle \mid \mathrm{CH}(W)_\alpha \neq 0\},$$

then

$$[\mathrm{CH}(W)] = \left[\sum_{\alpha \in I_W} \mathrm{CH}(X)_\alpha \right], \quad \text{where } I_W = \{\alpha \mid \mathrm{CH}(X)_\alpha \neq 0, \langle \alpha, \xi \rangle = \mathrm{CW}_\xi(X)\},$$

$$[\mathrm{CH}(Y)] = \left[\sum_{\alpha \in I_Y} \mathrm{CH}(W)_\alpha \right], \quad \text{where } I_Y = \{\alpha \mid \mathrm{CH}(W)_\alpha \neq 0, \langle \alpha, \eta \rangle = \mathrm{CW}_\eta(W)\}.$$

Note that $I_Y \subseteq I_W$. For any $\alpha \in M_{\mathbb{Z}}$ with $\mathrm{CH}(X)_\alpha \neq 0$, we have that $\langle \alpha, \xi \rangle \geq \mathrm{CW}_\xi(X)$, with equality if and only if $\alpha \in I_W$. Similarly, for any $\alpha \in M_{\mathbb{Z}}$ with $\mathrm{CH}(W)_\alpha \neq 0$ (and hence $\mathrm{CH}(X)_\alpha \neq 0$), we have that $\langle \alpha, \eta \rangle \geq \mathrm{CW}_\eta(W)$, with equality if and only if $\alpha \in I_Y$. So when $0 < \epsilon \ll 1$ and for any $\mathrm{CH}(X)_\alpha \neq 0$, we have that $\langle \alpha, \xi + \epsilon\eta \rangle \geq \mathrm{CW}_\xi(X) + \epsilon\mathrm{CW}_\eta(W)$, with equality if and only if $\alpha \in I_Y$. So we get

$$\lim_{t\to 0} \sigma_{\xi+\epsilon\eta}(t) \circ [\mathrm{CH}(X)] = \lim_{t\to 0} \left[\sum_\alpha t^{\langle \alpha, \xi+\epsilon\eta \rangle} \mathrm{CH}(X)_\alpha \right] = [\mathrm{CH}(Y)].$$

So for $0 < \epsilon \ll 1$, $\xi + \epsilon\eta$ induces an $\mathbb{R}$–test configuration that degenerates X to Y. By Lemma 2.11, we get the corresponding valuations v_ϵ.

Now we use the identity (113) to get

$$(170) \qquad \frac{d}{d\epsilon}\Big|_{\epsilon=0} \widetilde{\beta}(v_\epsilon) = \frac{d}{d\epsilon} H_Y^{\mathrm{NA}}(\mathcal{F}_{\mathrm{wt}_{\xi+\epsilon\eta}}) = \mathrm{Fut}_{Y,\xi}(\eta) < 0.$$

But this contradicts the assumption that $v_0 = v$ is the minimizer of $\widetilde{\beta}$.

Conversely, we need to show that a Ding-semistable valuation is a minimizer. Let $(\mathcal{X}, \mathcal{L}, a\eta)$ be any special test configuration of $(X, -K_X)$ and $\mathcal{F}_1 = \mathcal{F}_{(\mathcal{X}, \mathcal{L}, a\eta)}$ be the associated filtration. We consider the initial term degeneration of $\mathcal{F}_1$ with respect to $\mathcal{F}_0$ defined as above. Then we can use (166) to get

$$H_X^{\mathrm{NA}}(\mathcal{F}_1) \geq H_W^{\mathrm{NA}}(\mathcal{F}_1') \geq H_W^{\mathrm{NA}}(\mathcal{F}_{\mathrm{wt}_{\xi_0}}) = H^{\mathrm{NA}}(\mathcal{F}_0) = \tilde{\beta}(v),$$

where the second inequality follows from the results in Lemma 6.1 in the next section and the assumption that (W, ξ_0) is Ding-semistable. □

6 Uniqueness of minimizing special $\mathbb{R}$–test configurations

We prove Theorem 1.2 in this section. We first generalize the formula (113). Let $(X, -K_X, \mathbb{T}, \xi)$ be the data as before and $\mathcal{F}$ be a $\mathbb{T}$–equivariant filtration. We consider a family of $\mathbb{T}$–equivariant filtrations

$$(171) \qquad \mathcal{F}_s = s\mathcal{F}_{((1-s)/s)\xi} \quad \text{for } s \in (0, 1], \text{ with } \mathcal{F}_0 = \mathcal{F}_{\mathrm{wt}_\xi} \text{ and } \mathcal{F}_1 = \mathcal{F},$$

which interpolates $\mathcal{F}_{\mathrm{wt}_\xi}$ and $\mathcal{F}$.

Lemma 6.1 *For the family of filtrations* (171), *the following statements hold true*:

(i) *The map $s \mapsto H^{\mathrm{NA}}(\mathcal{F}_s)$ is smooth and convex. It is affine if and only if $G_\mathcal{F}$ is a multiple of $\langle x, \xi \rangle$.*

(ii) *We have the derivative formula*

$$(172) \qquad \frac{d}{ds}\Big|_{s=0} H^{\mathrm{NA}}(\mathcal{F}_s) = \beta_\xi(\mathcal{F}_{-\xi}).$$

To get (113) from (172), we just need to set $\mathcal{F} = \mathcal{F}_{\mathrm{wt}_{\xi+\eta}}$ so that $\mathcal{F}_s = \mathcal{F}_{\xi+s\eta}$. Moreover, we fix a faithful valuation that is adapted to the torus action (see Definition 2.21) and will freely use the associated Newton–Okounkov body $\Delta = \Delta(-K_X)$ of $(X, -K_X)$.

Proof By Lemma 2.22 and (25), as functions on $\Delta = \Delta(-K_X)$, we have

$$(173) \qquad G(s, y) := G_{\mathcal{F}_s}(y) = (1-s)\langle y, \xi \rangle + sG_\mathcal{F}(y).$$

So, by using Lemma 2.35, we get

$$(174) \qquad L^{\mathrm{NA}}(\mathcal{F}_s) = sL^{\mathrm{NA}}(\mathcal{F}),$$

$$(175) \qquad -\tilde{S}^{\mathrm{NA}}(\mathcal{F}_s) = \log\left(\frac{n!}{V}\int_\Delta e^{-G(s,y)}\, dy\right).$$

$L^{\mathrm{NA}}(\mathcal{F}_s)$ is linear in s and $-\tilde{S}^{\mathrm{NA}}(\mathcal{F}_s)$ is smooth in s. By Hölder's inequality, $-\tilde{S}^{\mathrm{NA}}(\mathcal{F}_s)$ is strictly convex in s unless $G_\mathcal{F}$ is a multiple of $\langle x, \xi \rangle$. This implies that $H^{\mathrm{NA}}(\mathcal{F}_s) = L^{\mathrm{NA}} - S^{\mathrm{NA}}$ is convex in $s \in [0, 1]$. To see (172), we calculate

$$\frac{d}{ds}\Big|_{s=0} H^{\mathrm{NA}}(\mathcal{F}_s) = L^{\mathrm{NA}}(\mathcal{F}) + \frac{\int_\Delta (\langle y, \xi \rangle - G_\mathcal{F}(y))e^{-G(0,y)}\, dx}{\int_\Delta e^{-G(0,y)}\, dy}$$

$$= L^{\mathrm{NA}}(\mathcal{F}_{-\xi}) - \frac{n!}{V_\xi}\int_\Delta G_{\mathcal{F}_{-\xi}}(y)e^{-\langle y, \xi \rangle}\, dy = \beta_\xi(\mathcal{F}_{-\xi}). \qquad \square$$

Assume that there are two special $\mathbb{R}$–test configurations $\mathcal{F}_i = \{\mathcal{F}_i R_m\}$ for $i = 0, 1$ of $(X, -K_X)$ that minimize $\boldsymbol{H}^{\mathrm{NA}}$. By Theorem 5.2, the central fibers $(W^{(i)} := \mathrm{Proj}(\mathrm{Gr}_{\mathcal{F}_i}), \xi_i = \xi_{\mathcal{F}_i})$ are both Ding-semistable. Now consider the initial term degeneration of $\mathcal{F}_1$ with respect to $\mathcal{F}_0$ as in the above section. We get a $\mathbb{T}_0$–equivariant filtration $\mathcal{F}_1'$ on $R' = R(W^{(0)}, -K_{W^{(0)}})$ and by (166), $\boldsymbol{H}_X^{\mathrm{NA}}(\mathcal{F}_1) \geq \boldsymbol{H}_{W^{(0)}}^{\mathrm{NA}}(\mathcal{F}_1')$.

Now, as at the beginning of this section, consider the family of filtrations that interpolates $\mathcal{F}_1'$ and $\mathcal{F}_{\mathrm{wt}_{\xi_0}} R' =: \mathcal{F}_{\mathrm{wt}_{\xi_0}}'$,

(176)
$$\mathcal{F}_s' := s\mathcal{F}_{((1-s)/s)\xi_0}' R'.$$

Applying Lemma 6.1 to $(W^{(0)}, \xi_0, \mathcal{F}_s')$, we know that $\boldsymbol{D}(s) := \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}_s')$ is convex in $s \in [0, 1]$. Moreover we have the relation

(177)
$$\boldsymbol{D}(0) = \boldsymbol{H}_{W^{(0)}}^{\mathrm{NA}}(\mathcal{F}_{\mathrm{wt}_{\xi_0}}) = \boldsymbol{H}_X^{\mathrm{NA}}(\mathcal{F}_0) = \boldsymbol{H}_X^{\mathrm{NA}}(\mathcal{F}_1) \geq \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}_1') = \boldsymbol{D}(1).$$

The 3$^{\mathrm{rd}}$ identity is by Theorem 5.2, that the $\mathcal{F}_i$ for $i = 0, 1$ both obtain the minimum of $\boldsymbol{H}^{\mathrm{NA}}$.

On the other hand, by (172),
$$\frac{d}{ds}\bigg|_{s=0} \boldsymbol{H}^{\mathrm{NA}}(\mathcal{F}_s') = \beta_{\xi_0}(\mathcal{F}_{-\xi_0}') \geq 0.$$

The last inequality is because $(W^{(0)}, \xi_0)$ is Ding-semistable.

By convexity of $\boldsymbol{D}(s)$, we conclude that $\boldsymbol{D}(s)$ is constant in s and by Lemma 6.1 that $G_{\mathcal{F}_1'}(y) \equiv \langle y, \xi_0 \rangle$ for any $y \in \Delta' = \Delta(W^{(0)}, -K_{W^{(0)}})$ (the Okounkov body of $(W^{(0)}, -K_{W^{(0)}})$).

By the discussion in previous section, we know that the relative successive minima of $\mathcal{F}_1$ with respect to $\mathcal{F}_0$ is the same as the relative successive minima of $\mathcal{F}_1'$ with respect to $\mathcal{F}_{\mathrm{wt}_{\xi_0}}'$, which is the same as the successive minima of $\mathcal{F}_{-\xi_0}'$ and is given by the difference $\lambda_k - \langle \alpha_k, \xi_0 \rangle$ with the notation there. So we get by Lemma 5.1 that

$$d_2(\mathcal{F}_0, \mathcal{F}_1)^2 = d_2(\mathcal{F}_0', \mathcal{F}_1')^2 = \lim_{m \to +\infty} \sum_k \frac{(\lambda_k - \langle \alpha_k, \xi_0 \rangle)^2}{m^2} = \lim_{m \to +\infty} \sum_i \frac{\lambda_i^{(m)}(\mathcal{F}_{-\xi_0}')^2}{m^2}$$

$$= \int_{\mathbb{R}} \lambda^2 \, \mathrm{DH}(\mathcal{F}_{-\xi_0}')^2 = \int_{\Delta'} G_{\mathcal{F}_{-\xi_0}'}^2 \, dy = \int_{\Delta'} (G_{\mathcal{F}'} - \langle y, \xi_0 \rangle)^2 \, dy = 0.$$

By [20], we know that $\mathcal{F}_0$ is asymptotically equivalent to $\mathcal{F}_1$. By Lemmas 2.11 and 2.29 (see also Proposition 2.27), we get $\mathcal{F}_0 = \mathcal{F}_1$.

Remark 6.2 Although here we are dealing with filtration of arbitrary ranks, the unique result in this section (and minimization result in previous section) can also be proved by using $r := \mathrm{rk}(\mathcal{F}_0)$–step degenerations to reduce to the rank-one case. To see this, we first choose $\{\eta_1, \ldots, \eta_r\} \in N_{\mathbb{Q}} \cong \mathbb{Q}^r$ (where $N = \mathrm{Hom}(\mathbb{C}^*, \mathbb{T}_0)$ as before) such that:

- $\mathrm{Span}_{\mathbb{R}}\{\eta_1, \ldots, \eta_r\} = N_{\mathbb{R}}$.
- For any $1 \leq k \leq r$, η_k induces a special test configuration whose central fiber is the same as $W^{(0)}$. This is achieved by choosing η_k satisfying $|\eta_k - \xi_0| \ll 1$.

By abuse of notation, we denote by $\mathcal{F}'_{\xi_0}$ (resp. $\mathcal{F}'_{\eta_1}$) the filtration on $R = R(X, -K_X)$ corresponding to the $\mathbb{R}$–test configuration induced by ξ_0 (resp. η_1), and also the filtration on $R' = R(W^{(0)}, -K_{W_0^{(0)}})$ corresponding to the weight filtration induced by ξ_0 (resp. η_k for $2 \le k \le r$). Set $\mathcal{F}_1^{\prime(0)} = \mathcal{F}_1$ and inductively define $\mathcal{F}_1^{\prime(k)}$ to be the initial term degeneration of $\mathcal{F}_1^{\prime(k-1)}$ with respect to $\mathcal{F}'_{\eta_k}$ for $1 \le k \le r$. By (166) for the rank-one case, we have

$$(178) \quad H_X^{\mathrm{NA}}(\mathcal{F}_1^{\prime(0)}) \ge H_{W^{(0)}}^{\mathrm{NA}}(\mathcal{F}_1^{\prime(1)}) \quad \text{and} \quad H_{W^{(0)}}^{\mathrm{NA}}(\mathcal{F}_1^{\prime(k-1)}) \ge H_{W^{(0)}}^{\mathrm{NA}}(\mathcal{F}_1^{\prime(k)}) \quad \text{for } 2 \le k \le r.$$

So if $\mathcal{F}_1 = \mathcal{F}_1^{\prime(0)}$ obtains the minimum of H_X^{NA}, then $\mathcal{F}^{\prime(k)}$ for any $1 \le k \le r$ also obtains the minimum of $H_{W^{(0)}}^{\mathrm{NA}}$. Now because $\mathcal{F}^{\prime(r)}$ is $\mathbb{T}_0$–invariant and $\mathcal{F}'_{\xi_0} = \mathcal{F}'_{\mathrm{wt}\xi_0}$ also obtains the minimum of $H_{W^{(0)}}^{\mathrm{NA}}$, we can use Lemma 2.38 to conclude that $\mathcal{F}^{\prime(r)} = \mathcal{F}'_{\xi_0}$.

On the other hand, by Lemma 5.1, we get for $2 \le k \le r$ that

$$d_2^X(\mathcal{F}_1^{\prime(0)}, \mathcal{F}'_{\eta_1}) = d_2^{W^{(0)}}(\mathcal{F}_1^{\prime(1)}, \mathcal{F}'_{\eta_1}) \quad \text{and} \quad d_2^{W^{(0)}}(\mathcal{F}_1^{\prime(k-1)}, \mathcal{F}'_{\eta_k}) = d_2^{W^{(0)}}(\mathcal{F}_1^{\prime(k)}, \mathcal{F}'_{\eta_k}).$$

So for any $1 \le k \le r$, we get, by omitting the superscripts and using the triangle inequality,

$$\begin{aligned}
d_2(\mathcal{F}_1^{\prime(k-1)}, \mathcal{F}'_{\xi_0}) &\le d_2(\mathcal{F}_1^{\prime(k-1)}, \mathcal{F}'_{\eta_k}) + d_2(\mathcal{F}'_{\eta_k}, \mathcal{F}'_{\xi_0}) \\
&= d_2(\mathcal{F}_1^{\prime(k)}, \mathcal{F}'_{\eta_k}) + d_2(\mathcal{F}'_{\eta_k}, \mathcal{F}'_{\xi_0}) \\
&\le d_2(\mathcal{F}_1^{\prime(k)}, \mathcal{F}'_{\xi_0}) + 2 d_2(\mathcal{F}'_{\eta_k}, \mathcal{F}'_{\xi_0}).
\end{aligned}$$

So we can inductively estimate

$$\begin{aligned}
d_2(\mathcal{F}_1, \mathcal{F}_0) = d_2(\mathcal{F}_1^{\prime(0)}, \mathcal{F}'_{\xi_0}) &\le d_2(\mathcal{F}_1^{\prime(1)}, \mathcal{F}'_{\xi_0}) + 2 d_2(\mathcal{F}'_{\eta_1}, \mathcal{F}'_{\xi_0}) \\
&\le d_2(\mathcal{F}_1^{\prime(2)}, \mathcal{F}'_{\xi_0}) + 2(d_2(\mathcal{F}'_{\eta_2}, \mathcal{F}'_{\xi_0}) + d_2(\mathcal{F}'_{\eta_1}, \mathcal{F}'_{\xi_0})) \\
&\vdots \\
&\le d_2(\mathcal{F}_1^{\prime(r)}, \mathcal{F}'_{\xi_0}) + 2 \sum_{k=1}^{r} d_2(\mathcal{F}'_{\eta_k}, \mathcal{F}'_{\xi_0}) \\
&= 2 \sum_{k=1}^{r} d_2(\mathcal{F}'_{\eta_k}, \mathcal{F}'_{\xi_0}).
\end{aligned}$$

Now we can choose η_k so that $d_2(\mathcal{F}'_{\eta_k}, \mathcal{F}'_{\xi_0})$ is arbitrarily small for all $1 \le k \le r$. So we indeed get $d_2(\mathcal{F}_1, \mathcal{F}_0) = 0$, as desired.

7 Cone construction and g–normalized volume

Let X be an n–dimensional $\mathbb{Q}$–Fano variety and for simplicity of notation, assume that $-K_X$ is Cartier. Recall that $R = \bigoplus_m R_m = \bigoplus H^0(X, m(-K_X))$. We define the cone

$$(179) \quad\quad\quad C = C(X, -K_X) = \mathrm{Spec}_{\mathbb{C}} R, \quad o = \mathfrak{m} = \bigoplus_{m>0} R_m.$$

Then (C, o) is a klt cone singularity.

Since X admits a $\mathbb{C}^* \times \mathbb{T}$–action, we have a decomposition of the coordinate ring of R,

$$(180) \qquad R = \bigoplus_{m \geq 0} \bigoplus_{\alpha \in \mathbb{Z}^r} R_{m,\alpha}.$$

For any $\mathbb{T}$–invariant homogeneous primary ideal $\mathfrak{a} = \bigoplus_m \bigoplus_\alpha \mathfrak{a}_{m,\alpha} \subset R$, define the g–colength and g–multiplicity of $\mathfrak{a}$ by

$$(181) \qquad \mathrm{colen}_g(\mathfrak{a}) = \sum_{m \geq 0} \sum_\alpha g\left(\frac{\alpha}{m}\right) \dim R_{m,\alpha}/\mathfrak{a}_{m,\alpha},$$

$$(182) \qquad \mathrm{mult}_g(\mathfrak{a}) = \lim_{k \to +\infty} \frac{\mathrm{colen}_g(\mathfrak{a}^k)}{k^{n+1}/(n+1)!}.$$

See [63] for the study of such equivariant multiplicity. More generally, let $\mathfrak{a}_\bullet = \{\mathfrak{a}_k\}_{k \in \mathbb{N}}$ be a graded sequence of $\mathbb{C}^* \times \mathbb{T}$–invariant ideals. We define

$$(183) \qquad \mathrm{mult}_g(\mathfrak{a}_\bullet) = \lim_{k \to +\infty} \frac{\mathrm{colen}_g(\mathfrak{a}_k)}{k^{n+1}/(n+1)!}.$$

One can use the techniques of Newton–Okounkov bodies to show that the limit exists. To see this, we can adapt the argument in the work in [46] as follows. First choose a valuation $\mathfrak{v}$ adapted to the $\mathbb{T}$–action on X (in the sense of Definition 2.21). We can construct a $\mathbb{C}^* \times \mathbb{T}$–invariant $\mathbb{Z}^{n+1}$–valuation on C by

$$(184) \qquad \mathfrak{V}(f) = (m, \mathfrak{v}(f)) \quad \text{for any } f \in R_m.$$

Denote by $\mathfrak{C}$ the strongly convex cone which is the closure of the convex hull of the value semigroup $\mathfrak{V}(R)$. To each graded sequence of $\mathbb{C}^* \times \mathbb{T}$–invariant ideals $\mathfrak{a}_\bullet$, one can associate a convex region $\bar{P} := \bar{P}(\mathfrak{a}_\bullet) \subset \mathfrak{C}$ such that $\bar{P}^c := \mathfrak{C} \setminus \bar{P}$ is bounded. If we still denote by $g(y)$ the pullback of function g by the projection $\mathbb{R}^{n+1} = \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, then mult_g is given by the weighted volume of the co-convex set $\bar{P}^c$,

$$(185) \qquad \mathrm{mult}_g(\mathfrak{a}_\bullet) = (n+1)! \int_{\bar{P}^c} g(y) \, dy.$$

Let $\mathrm{Val}_{C,o}$ be the space of real valuations that are centered at o, and by $\mathrm{Val}_{C,o}^{\mathbb{C}^* \times \mathbb{T}}$ the subset of $\mathbb{C}^* \times \mathbb{T}$–invariant real valuations in $\mathrm{Val}_{C,o}$. If $\widetilde{C} \to C$ is the blowup of the vertex $o \in X$, then the exceptional divisor on $\widetilde{C}$ is isomorphic to X, and we will denote by ord_X the associated divisorial valuation contained in $\mathrm{Val}_{C,o}^{\mathbb{C}^* \times \mathbb{T}}$.

Let $\bar{v} \in \mathrm{Val}_{C,o}^{\mathbb{C}^* \times \mathbb{T}}$ be any $\mathbb{C}^* \times \mathbb{T}$–invariant valuation. Then for any $\lambda \in \mathbb{R}$, $\mathfrak{a}_\lambda(\bar{v}) = \{f \in R \mid \bar{v}(f) > m\}$ is a $\mathbb{T}$–invariant homogeneous primary ideal. Set $\mathfrak{a}_\bullet(\bar{v}) = \mathfrak{a}_m(\bar{v})$ and define (see [35] for the $g = 1$ case)

$$\mathrm{vol}_g(\bar{v}) := \mathrm{mult}_g(\mathfrak{a}_\bullet(v)) = \lim_{m \to +\infty} \frac{\mathrm{colen}_g(\mathfrak{a}_\lambda(\bar{v}))}{\lambda^{n+1}/(n+1)!}.$$

We define the following equivariant version of normalized volume [51]:

$$\widehat{\mathrm{vol}}_g : \mathrm{Val}_{C,o}^{\mathbb{C}^* \times \mathbb{T}} \to \mathbb{R}_{>0} \cup \{+\infty\}, \qquad \widehat{\mathrm{vol}}_g(\bar{v}) = \begin{cases} A_C(\bar{v})^{n+1} \cdot \mathrm{vol}_g(\bar{v}) & \text{when } A_C(\bar{v}) < +\infty, \\ +\infty & \text{otherwise.} \end{cases}$$

By using the same argument as in the study of normalized volumes, one can generalize almost all the results about normalized volume to work for the g–normalized volume functional. Here we just write down a few results that we need in the next section. We have the following equivariant version of an identity from [60].

Lemma 7.1 *With the above notation, we have the identity*

$$\inf_{\overline{v}} \widehat{\mathrm{vol}}_g(\overline{v}) = \inf_{\mathfrak{a}} \mathrm{lct}(\mathfrak{a})^n \cdot \mathrm{mult}_g(\mathfrak{a}) = \inf_{\mathfrak{a}_\bullet} \mathrm{lct}(\mathfrak{a}_\bullet)^n \cdot \mathrm{mult}_g(\mathfrak{a}_\bullet), \tag{186}$$

where $\overline{v}$ ranges over $\mathbb{C}^* \times \mathbb{T}$–invariant valuations, and $\mathfrak{a}$ (resp. $\mathfrak{a}_\bullet$) ranges over $\mathbb{C}^* \times \mathbb{T}$–invariant ideals (resp. graded sequences of $\mathbb{C}^* \times \mathbb{T}$–invariant ideals).

This is proved by using exactly the same argument. For the reader's convenience, we give the short proof.

Proof For any $\overline{v} \in \overset{\circ}{\mathrm{Val}}_{C,o}$, we have

$$\mathrm{lct}(\mathfrak{a}_\bullet(\overline{v}))^n \cdot \mathrm{mult}_g(\mathfrak{a}_\bullet(\overline{v})) \le \left(\frac{A_C(\overline{v})}{\overline{v}(\mathfrak{a}_\bullet)} \right)^n \mathrm{vol}_g(\overline{v}) = A_C(\overline{v})^n \mathrm{vol}_g(\overline{v}). \tag{187}$$

Conversely, for any graded sequence of ideals $\mathfrak{a}_\bullet$, let $\overline{w} \in \overset{\circ}{\mathrm{Val}}_{C,o}$ be the valuation that calculates $\mathrm{lct}(\mathfrak{a}_\bullet)$, which exists by [44]. By multiplying by a constant, we can assume $1 = \overline{w}(\mathfrak{a}_\bullet) = \inf_m \overline{w}(\mathfrak{a}_m)/m$. So $\mathfrak{a}_m \subseteq \mathfrak{a}_m(\overline{w})$, which implies $\mathrm{mult}_g(\mathfrak{a}_\bullet) \ge \mathrm{mult}_g(\mathfrak{a}_\bullet(\overline{w})) = \mathrm{vol}_g(\overline{w})$. Then we get

$$\mathrm{lct}(\mathfrak{a}_\bullet)^n \cdot \mathrm{mult}_g(\mathfrak{a}_\bullet) = \left(\frac{A_C(\overline{w})}{\overline{w}(\mathfrak{a}_\bullet)} \right)^n \cdot \mathrm{mult}_g(\mathfrak{a}_\bullet) \ge A_C(\overline{w})^n \cdot \mathrm{vol}_g(\overline{w}) = \widehat{\mathrm{vol}}_g(\overline{w}). \qquad \square \tag{188}$$

For any $v \in X_{\mathbb{Q}}^{\mathrm{div}}$ and $\tau > 0$, we denote by $\overline{v}_\tau$ the $\mathbb{C}^*$–invariant valuation on C given by

$$\overline{v}_\tau \left(\sum_i f_i t^i \right) = \min_i (v(f_i) + \tau i). \tag{189}$$

By using the same calculation as in [50], we get:

Theorem 7.2 *The g–volume of $\overline{v}_\tau$ is given by the formula*

$$\mathrm{vol}_g(\overline{v}_\tau) = \frac{1}{\tau^{n+1}} V_g - (n+1) \int_0^{+\infty} \mathrm{vol}_g(\mathcal{F}_v R^{(x)}) \frac{dx}{(x+\tau)^{n+2}}. \tag{190}$$

We have the following criterion for g–Ding-semistability, which generalizes the results in [50; 53; 58] about normalized volumes.

Theorem 7.3 *The pair (X, η) is g–Ding-semistable if and only if ord_X obtains the minimum of $\widehat{\mathrm{vol}}_g$ over $\mathrm{Val}_{C,o}^{\mathbb{C}^* \times \mathbb{T}}$.*

Proof For any $v \in (X_{\mathbb{Q}}^{\mathrm{div}})^{\mathbb{T}}$, consider $w_s := \overline{(sv)}_{(1-s)A_X(v)} \in \mathrm{Val}_{C,o}^{\mathbb{C}^* \times \mathbb{T}}$. Then $w_0 = A_X(v)\overline{v}_0$ and $w_1 = v$. We also have $A_C(w_s) \equiv A_X(v)$. Set

$$f(s) = \widehat{\mathrm{vol}}(w_s) = A_C(w_s)^{n+1} \mathrm{vol}_g(w_s)$$

$$= A_X(v)^{n+1} \left(\frac{V_g}{(1-s)^{n+1} A_X(v)^{n+1}} - (n+1) \int_0^{+\infty} \mathrm{vol}_g(\mathcal{F}_v R^{(x)}) \frac{s\, dx}{(sx + (1-s)A_X(v))^{n+2}} \right)$$

$$= A_X(v)^{n+1} \int_0^{+\infty} \frac{-d\,\mathrm{vol}_g(\mathcal{F}_v R^{(x)})}{(sx + (1-s)A_X(v))^{n+1}}.$$

Then $f(s)$ is a convex function in $s \in [0,1]$. Its derivative at $s = 0$ is given by

$$(191) \qquad f'(0) = A_X(v)^{n+1} \left((n+1) \frac{V_g}{A_X(v)^{n+1}} - (n+1) \int_0^{+\infty} \mathrm{vol}_g(\mathcal{F}_v R^{(x)})\, dx\, \frac{1}{A_X(v)^{n+2}} \right)$$

$$= \frac{n+1}{A_X(v)V_g} \left(A_X(v) - \frac{1}{V_g} \int_0^{+\infty} \mathrm{vol}_g(\mathcal{F}_v R^{(x)})\, dx \right)$$

$$= \frac{n+1}{A_X(v)V_g} \cdot \beta_g(v).$$

With this and Theorem 2.47, we can easily derive the conclusion as in [50]. $\qquad\square$

Remark 7.4 By the same argument as in the case of normalized volume [12; 78], one shows that g–Ding-semistability is Zariski-open for a $\mathbb{T}$–equivariant family of Fano varieties.

8 Uniqueness of polystable degeneration

In this section, we prove Theorem 1.3. The proof is verbatim the same as the proof of the existence and uniqueness of K–polystable degenerations for any K–semistable $\mathbb{Q}$–Fano varieties, as proved in [55]; see also [16]. Indeed, we just need to carry out the same argument by using the equivariant version of normalized volume and the modified Futaki invariant Fut_ξ, etc. To avoid redundancy, we only sketch the key steps and refer to [55; 16] for more details.

Assume that (X, ξ) is semistable and admits two polystable degenerations via two special test configurations $(\mathcal{X}^{(i)}, -K_{\mathcal{X}^{(i)}})$ for $i = 0, 1$. Take cones fiberwise to get a special test configuration of Fano cones $(C^{(i)}), \zeta^{(i)})$, where $\zeta^{(i)}$ is the radial vector field.

Let E_k be the Kollár component (see [58] for the definition) obtained by blowing up the vertex of $C^{(0)}$ with weight $(k, 1)$. Then we have

$$\widehat{\mathrm{vol}}_g(E_k) = \widehat{\mathrm{vol}}_g(\mathrm{ord}_X) + O(k^{-2}).$$

Set $\mathfrak{a}_\bullet = \{\mathfrak{a}_\ell(E_k)\}$. Then

$$\mathrm{lct}(\mathfrak{a}_\bullet) = \frac{A(E_k)}{\mathrm{ord}_{E_k}(\mathfrak{a}_\bullet)} = A(E_k) =: c_k = O(k), \quad \mathrm{lct}(X, \mathfrak{a}_\bullet)^n \cdot \mathrm{mult}_g(\mathfrak{a}_\bullet) = \widehat{\mathrm{vol}}_g(E_k).$$

Consider the initial degeneration of $\mathfrak{a}_\bullet$ with respect to $\mathcal{C}^{(1)}$,

$$(192) \qquad \mathbf{in}(\mathfrak{a}_\ell) = \mathrm{span}_{\mathbb{C}}\{\mathbf{in}(f), f \in \mathfrak{a}_\ell(E_k)\}.$$

Using the preservation of co-length under initial term degeneration, we get

$$
\begin{aligned}
\mathrm{lct}^n(\mathcal{C}_0^{(1)}, \mathbf{in}(\mathfrak{a}_\bullet)) &\geq \frac{\widehat{\mathrm{vol}}_g(\mathrm{ord}_{\mathcal{X}_0^{(1)}})}{\mathrm{mult}_g(\mathbf{in}(\mathfrak{a}_\bullet))} = \frac{V_g}{\mathrm{mult}_g(\mathfrak{a}_\bullet)} \\
&= \frac{V_g}{\widehat{\mathrm{vol}}_g(E_k)}\mathrm{lct}(\mathfrak{a}_\bullet)^n = \frac{V_g}{V_g + O(k^{-2})}\mathrm{lct}(\mathfrak{a}_\bullet)^n \\
&= (1 + O(k^{-2}))c_k = c_k + O(k^{-1}).
\end{aligned}
$$

Let $Z_k \to \mathcal{C}^{(0)}$ be the extraction of E_k, and let $Z_k \times \mathbb{C}^*$ be the product along $\mathcal{C}^{(1)} \setminus \mathcal{C}_0^{(1)} \cong C \times \mathbb{C}^*$ with exceptional divisor $\mathcal{E}_k$. Let $\mathfrak{B}_\bullet = \{\mathfrak{B}_\ell\}$ be ideal on the total space $\mathcal{C}^{(1)}$ obtained by the above degenerating $\mathfrak{a}_\ell$. Then we have

$$(193) \qquad A(\mathcal{C}^{(1)}, c_k(1 - \epsilon k^{-1})\mathfrak{B}_\bullet, \mathcal{E}_k) = A(C, c_k(1 - \epsilon k^{-1})\mathfrak{a}_\bullet, E_k) = \epsilon k^{-1}c_k = \epsilon O(1),$$

$$(194) \qquad \mathrm{lct}(\mathcal{C}_0^{(1)}, c_k(1 - \epsilon k^{-1})\mathbf{in}(\mathfrak{a}_\bullet)) \geq c_k^{-1}(1 - \epsilon k^{-1})(c_k + O(k^{-1})) = 1 - \epsilon k^{-1} + O(k^{-2}).$$

By inversion of adjunction,

$$(195) \qquad \mathrm{lct}(\mathcal{C}^{(1)}, c_k(1 - \epsilon k^{-1})\mathfrak{B}_\bullet) \geq 1 - \epsilon k^{-1} + O(k^{-2}).$$

When $0 < \epsilon \ll 1$, by [9], we can extract the divisor $\mathcal{E}_k$ over $\mathcal{C}^{(1)}$. By the same argument as [55], we get the commutative diagram

(196)

By the same argument as in [55], we know that both test configurations $\mathcal{X}'^{(i)}$ for $i = 0, 1$ are weakly special and have vanishing Fut_ξ invariant. By [39], we know that both of them are special and hence $X_0^{(1)} \cong X_0' \cong X_0^{(0)}$ by the polystability of $X_0^{(i)}$.

The existence part can again be proved by the similar arguments as in [55], which deals with the case when $\xi = 0$. We just sketch the arguments. If (X, ξ) is K–polystable, then we are done. Otherwise, we can find a nontrivial $\mathbb{T}$–equivariant special test configuration such that the central fiber (with the vector

field ξ) has a vanishing Fut_ξ invariant. By [55, Proof of Lemma 3.1], we know that the central fiber is K–semistable, and has an effective action by a larger torus. If the central fiber is K–polystable, then we are done again. Otherwise, we can continue this process, which must stop since the dimension of the torus is bounded by the dimension of X.

Proof of Corollary 1.4 By the work of Chen, Sun and Wang in [28], which is based on the resolution of the Hamilton–Tian conjecture [29], we get a special $\mathbb{R}$–test configuration $\mathcal{F}^{\mathrm{ss}}$ with central fiber (W, ξ), and a special test configuration of (W, ξ) with central fiber (X_∞, ξ), which admits a Kähler–Ricci soliton and hence is K–polystable. By the work of Dervan and Székelyhidi [31], $\mathcal{F}^{\mathrm{ss}}$ obtains the minimum $h(X)$. The statement follows directly from Theorems 1.2 and 1.3. □

Remark 8.1 The fact that $\mathcal{F}^{\mathrm{ss}}$ obtains the minimum also follows from the K–semistability of (W, ξ) and Theorem 5.2. The K–semistability of (W, ξ) follows from the same degeneration argument as used in [58], or the Zariski-openness of K–semistability as pointed out in Remark 7.4.

Remark 8.2 As in the more general setting of [55] or [52], the algebraic results in this paper can be generalized to the log Fano case in a straightforward way.

Appendix   Properties of $\widetilde{S}(v)$

Recall that by (137), for any valuation $v \in \overset{\circ}{\mathrm{Val}}(X)$ we have

$$(197) \qquad Q(v) := Q(\mathcal{F}_v) = e^{-\widetilde{S}^{\mathrm{NA}}(\mathcal{F}_v)} = 1 - \frac{1}{V} \int_0^{T(v)} e^{-x} \mathrm{vol}(\mathcal{F}_v^{(x)} R_\bullet) \, dx =: 1 - \Psi(v),$$

where, for simplicity of notation, we have written

$$(198) \qquad T(v) = \lambda_{\max}(\mathcal{F}_v) \quad \text{and} \quad \Psi(v) = \frac{1}{V} \int_0^{T(v)} e^{-x} \mathrm{vol}(\mathcal{F}_v^{(x)} R_\bullet) \, dx.$$

Proposition A.1 *The function $v \mapsto \Psi(v)$ is strictly increasing on $\overset{\circ}{\mathrm{Val}}(X)$. In other words, if $v \leq w$, then $\Psi(v) \leq \Psi(w)$, with the identity true only if $v = w$. As a consequence, $v \mapsto \widetilde{S}(v)$ is strictly increasing on* $\mathrm{Val}(X)$.

This is proved as in [11, Proof of Proposition 3.15] (which is based on an argument in the local case from [58]). We sketch the argument for the reader's convenience.

Proof First, by using Theorem 2.5, we can show that

$$(199) \qquad \Psi(v) = \lim_{m \to +\infty} \frac{1}{m N_m} \sum_{j \geq 1} e^{-j/m} \dim \mathcal{F}_v^j R_m.$$

Suppose that $v \leq w$ but $v \neq w$. Then by rescaling v, w and $L = -K_X$, we can assume that there exists $s \in H^0(X, L)$ with $w(s) = p \in \mathbb{N}^*$ and $v(s) \leq p - 1$. Then, arguing as in [11, Proof of Proposition 3.15],

we have

$$(200) \qquad \dim(\mathcal{F}_w^j R_m / \mathcal{F}_v^j R_m) \geq \sum_{1 \leq i \leq \min\{j/p, m\}} \dim(\mathcal{F}_v^{j-ip} R_m / \mathcal{F}_v^{j-ip+1} R_{m-i}).$$

One the other hand, with $C = \max\{T(v), T(w)\}$, we get

$$\sum_{j \geq 1} \dim e^{-j/m}(\mathcal{F}_w^j R_m - \mathcal{F}_v^j R_m) \geq e^{-C} \sum_{j \geq 1}(\mathcal{F}_w^j R_m - \mathcal{F}_v^j R_m)$$

$$\geq e^{-C} \sum_{1 \leq i \leq m} \sum_{j \geq pi}(\dim \mathcal{F}_v^{j-ip} R_{m-i} - \mathcal{F}_v^{j-ip+1} R_{m-i})$$

$$= e^{-C} \sum_{1 \leq i \leq m} \dim R_{m-i}.$$

So we conclude

$$\Psi(v) - \Psi(w) \geq e^{-C} \lim_{m \to +\infty} \frac{1}{mN_m} \sum_{1 \leq i \leq m} \dim R_{m-i} > 0. \qquad \square$$

Let $\pi\colon Y \to X$ be a proper birational morphism with Y a regular and $E = \sum_i E_i$ a reduced simple normal crossing divisor.

Proposition A.2 *The function $v \mapsto \boldsymbol{Q}(v)$ is continuous on* $\mathrm{QM}(Y, E)$.

We use the same strategy as [14, Proposition 2.4]. As noted in [38], for any $v \in \mathrm{Val}(X)$, we have $A(v)/T(v) \geq \alpha(X) > 0$, which implies, with $C = \alpha(X)^{-1}$,

$$(201) \qquad\qquad\qquad\qquad\qquad\qquad T(v) \leq CA(v).$$

Lemma A.3 *For any $v \in \mathrm{Val}(X)$, we have the inequality*

$$(202) \qquad\qquad\qquad\qquad\qquad\qquad \Psi(v) \leq CA(v).$$

Proof Since $\mathrm{vol}(\mathcal{F}^{(x)} R_\bullet) \leq V$, we immediately get

$$\Psi(v) \leq \int_0^{T(v)} e^{-x} \, dx = 1 - e^{-T(v)} \leq T(v) \leq CA(v),$$

where we used the inequality $1 - e^{-x} \leq x$ for any $x \in \mathbb{R}_{\geq 0}$, and the inequality (201). $\qquad \square$

Similarly to [38; 11], we introduce the approximation

$$(203) \qquad \boldsymbol{Q}_m(\mathcal{F}) = \frac{1}{N_m} \sum_i e^{-\lambda_i^{(m)}/m} = \frac{1}{N_m} \int_0^{+\infty} e^{-x/m} d(-\dim \mathcal{F}^x R_m)$$

$$(204) \qquad\qquad = 1 - \frac{1}{N_m} \int_0^{\lambda_{\max}^{(m)}(\mathcal{F})/m} e^{-x} \dim \mathcal{F}^{xm} R_m \, dx =: 1 - \Psi_m(\mathcal{F}),$$

where we set

$$(205) \qquad \Psi_m(v) = \frac{1}{N_m} \int_0^{\lambda_{\max}^{(m)}} e^{-x} \dim \mathcal{F}^{xm} R_m \, dx = \frac{1}{N_m} \int_0^{T(v)} e^{-x} \dim \mathcal{F}^{xm} R_m \, dx.$$

Similarly to [38; 11], for any valuation $v \in \mathrm{Val}(X)$ we have the identity

(206)
$$\boldsymbol{Q}_m(v) = \boldsymbol{Q}_m(\mathcal{F}_v) = \min_{\{s_j\}} \frac{1}{N_m} \sum_{j=1}^{N_m} e^{-v(s_j)/m},$$

where the minimum is taken over all bases $s_1, \ldots, s_{N_m}$ of $H^0(X, -mK_X)$.

For any $\mathfrak{s} := \{s_1, \ldots, s_{N_m}\} \in H^0(X, -mK_X)^{N_m}$, define a function

(207)
$$\varphi_{\mathfrak{s}}(v) := \sum_{j=1}^{N_m} e^{-v(s_j)/m}.$$

By the same argument as in [14, Proof of Lemma 2.5], the set of functions $\{\varphi_{\mathfrak{s}}(v) \mid \mathfrak{s} \in R_m^{N_m}\}$ is finite. So $\boldsymbol{Q}_m$ is continuous on $\mathrm{QM}(Y, E)$.

As in [14, Proof of Proposition 2.4], the continuity of Ψ and hence $\boldsymbol{Q}$ follows easily from the following proposition, which we prove by using the techniques developed in [11; 13].

Lemma A.4   (i) *For any $v \in \mathrm{Val}(X)$ with $A(v) < +\infty$, we have the convergence*

(208)
$$\lim_{m \to +\infty} \Psi_m(v) = \Psi(v).$$

  (ii) *For any $\epsilon > 0$ and any $C_1 > 0$, there exists $C_2 > 0$ and $m_0 > 0$ such that if $v \in \mathrm{Val}(X)$ satisfies $A(v) < C_1$, we have*

(209)
$$|\Psi_m(\mathcal{F}_v) - \Psi(\mathcal{F}_v)| \leq \epsilon$$

  for all m divisible by m_0.

Proof The first statement follows from Theorem 2.5(ii). We focus on the second statement.

Note that e^{-G} is convex and $0 \leq e^{-G} \leq 1$. By [11, Lemma 2.2], for any $\epsilon' > 0$ there exists $m_0(\epsilon')$ such that for any $m \geq m_0$,

(210)
$$\int_\Delta e^{-G} \, d\rho_m \geq \int_\Delta e^{-G} \, dy - \epsilon'.$$

By the same argument as [11, Proof of Lemma 2.9], we get

(211)
$$\boldsymbol{Q}_m(\mathcal{F}_v) \geq \frac{m^n}{N_m} \int_\Delta e^{-G} \, d\rho_m.$$

Note that $\lim_{m \to +\infty} m^n / N_m = V$. So for any $\epsilon > 0$ there exists m_0 such that for any $m \geq m_0$,

(212)
$$\boldsymbol{Q}_m(\mathcal{F}_v) \geq \frac{n!}{V} \int_\Delta e^{-G} \, dy - \epsilon = \boldsymbol{Q}(\mathcal{F}_v) - \epsilon.$$

We need to prove the other direction of inequality. Following [11], define a graded linear series

(213)
$$\widetilde{\mathcal{F}}_{m,p}^{(t)} R_{mp} := H^0\big(X, mpL \otimes \overline{\mathfrak{b}(|\mathcal{F}^{mt} R_m|)^p}\big),$$

where $\mathfrak{b}(|\mathcal{F}^{mt} R_m|)$ is the base ideal of the sublinear system $\mathcal{F}^{tm} R_m$. Set

$$(214) \qquad \widetilde{\Psi}_m(\mathcal{F}) = \int_0^{T(v)} e^{-t} \operatorname{vol}(\widetilde{\mathcal{F}}_{m,\bullet}^{(t)}) \, dt.$$

By [11, Proposition 5.13], there exists $a = a(X, -K_X) > 0$ such that for all $t \in \mathbb{Q}_{>0}$ with $mt > A(v)$, we have, with $t' = t - m^{-1}(at + A(v))$,

$$(215) \qquad \left(\frac{m-a}{m}\right)^{n+1} \operatorname{vol}(\mathcal{F}_\bullet^{(t)}) \leq \frac{1}{m^n} \operatorname{vol}(\widetilde{\mathcal{F}}_{m,\bullet}^{(t')}).$$

So we can estimate as in [11, Proof of Proposition 5.15] to get

$$\widetilde{\Psi}_m(v) \geq \left(\frac{m-a}{m}\right)^{n+1} \left(\Psi(v) - e^{(aT(v)+A(v))/m} \int_0^{A(v)/(m-a)} \frac{\operatorname{vol}(\mathcal{F}^{(t)})}{V} e^{-t} \, dt \right)$$

$$\geq \left(\frac{m-a}{m}\right)^{n+1} \left(\Psi(v) - e^{CA(v)/m} \frac{A(v)}{m-a} \right).$$

From this it is easy to get

$$\Psi(v) - \widetilde{\Psi}_m(v) \leq C \frac{A(v)}{m}.$$

To compare with Ψ_m, we further set

$$(216) \qquad \mathcal{F}_{m,p}^{(x)} = \operatorname{Im}\left(S^p \mathcal{F}^{mx} R_m \to H^0(X, pmL) \right).$$

By [13, Propositions 5.14 and 3.2], there exists a positive constant $C > 0$ independent of v such that for all $x \leq T(v) - CA(v)/m$, we have $\operatorname{vol}(\mathcal{F}_{m,\bullet}^{(x)}) = \operatorname{vol}(\widetilde{\mathcal{F}}_{m,\bullet}^{(x)})$. So as in [13, Proof of Proposition 5.15], we get

$$\Psi(v) \leq \widetilde{\Psi}_m(v) + C \frac{A(v)}{m} = \frac{1}{V} \int_0^{T(v)} \frac{\operatorname{vol}(\widetilde{\mathcal{F}}_{m,\bullet}^{(x)})}{m^n} e^{-x} \, dx + \frac{CA(v)}{m}$$

$$\leq \frac{1}{V} \int_0^{T(v)-CA(v)/m} \frac{\operatorname{vol}(\mathcal{F}_{m,\bullet}^{(x)})}{m^n} e^{-x} \, dx + \frac{CA(v)}{m}$$

$$\leq \frac{1}{V} \int_0^{T(v)} \frac{\operatorname{vol}(\mathcal{F}_{m,\bullet}^{(x)})}{m^n} e^{-x} \, dx + \frac{CA(v)}{m}.$$

For the second inequality we used the estimate that, as $m \to +\infty$,

$$\int_{T(v)-CA(v)/m}^{T(v)} e^{-x} \, dx = e^{-(T(v)-CA(v)/m)} - e^{-T(v)} \leq e^{CA(v)/m} - 1 = O\left(\frac{A(v)}{m}\right).$$

Fixing any $\epsilon > 0$, by choosing $m \gg 1$ and $p \gg 1$ we have (see [13, equation (5.6)]):

$$(217) \qquad \left| \frac{\operatorname{vol}(\mathcal{F}_{m,\bullet}^{(x)})}{m^n V} - \frac{\dim \mathcal{F}_{m,p}^{(x)}}{N_{mp}} \right| < \epsilon.$$

Finally we can estimate as in [13, Proof of Theorem 5.13]: for $m \gg 1$,

$$\Psi(v) \leq \frac{1}{V} \int_0^{T(v)} \frac{\operatorname{vol}(\mathcal{F}_{m,\bullet}^{(x)})}{m^n} e^{-x} \, dx + \frac{CA(v)}{m} \leq \int_0^{+\infty} \frac{\dim \mathcal{F}_{m,p}^{(x)}}{N_{mp}} e^{-x} \, dx + \epsilon T(v) + \frac{CA(v)}{m}$$

$$\leq \int_0^{+\infty} \frac{\dim \mathcal{F}^{pmx} R_m}{N_{mp}} e^{-x} \, dx + 2\epsilon A(v) = \Psi(v) + 2\epsilon A(v).$$

In the third inequality, we used again the inequality $1 - e^{-T(v)} \leq T(v)$. Since $\epsilon > 0$ is arbitrary, we get the conclusion. $\qquad\square$

References

[1] **H Abban**, **Z Zhuang**, *K–stability of Fano varieties via admissible flags*, Forum Math. Pi 10 (2022) art. id. e15 MR Zbl

[2] **K Altmann**, **J Hausen**, *Polyhedral divisors and algebraic torus actions*, Math. Ann. 334 (2006) 557–607 MR Zbl

[3] **K Altmann**, **J Hausen**, **H Süss**, *Gluing affine torus actions via divisorial fans*, Transform. Groups 13 (2008) 215–242 MR Zbl

[4] **R Bamler**, *Convergence of Ricci flows with bounded scalar curvature*, Ann. of Math. 188 (2018) 753–831 MR Zbl

[5] **R J Berman**, *K–polystability of $\mathbb{Q}$–Fano varieties admitting Kähler–Einstein metrics*, Invent. Math. 203 (2016) 973–1025 MR Zbl

[6] **R J Berman**, **S Boucksom**, **P Eyssidieux**, **V Guedj**, **A Zeriahi**, *Kähler–Einstein metrics and the Kähler–Ricci flow on log Fano varieties*, J. Reine Angew. Math. 751 (2019) 27–89 MR Zbl

[7] **R J Berman**, **S Boucksom**, **M Jonsson**, *A variational approach to the Yau–Tian–Donaldson conjecture*, J. Amer. Math. Soc. 34 (2021) 605–652 MR Zbl

[8] **R J Berman**, **D W Nystrom**, *Complex optimal transport and the pluripotential theory of Kähler–Ricci solitons*, preprint (2014) arXiv 1401.8264

[9] **C Birkar**, **P Cascini**, **C D Hacon**, **J McKernan**, *Existence of minimal models for varieties of log general type*, J. Amer. Math. Soc. 23 (2010) 405–468 MR Zbl

[10] **H Blum**, *Existence of valuations with smallest normalized volume*, Compos. Math. 154 (2018) 820–849 MR Zbl

[11] **H Blum**, **M Jonsson**, *Thresholds, valuations, and K–stability*, Adv. Math. 365 (2020) art. id. 107062 MR Zbl

[12] **H Blum**, **Y Liu**, *The normalized volume of a singularity is lower semicontinuous*, J. Eur. Math. Soc. 23 (2021) 1225–1256 MR Zbl

[13] **H Blum**, **Y Liu**, *Openness of uniform K–stability in families of $\mathbb{Q}$–Fano varieties*, Ann. Sci. Éc. Norm. Supér. 55 (2022) 1–41 MR Zbl

[14] **H Blum**, **Y Liu**, **C Xu**, *Openness of K–semistability for Fano varieties*, Duke Math. J. 171 (2022) 2753–2797 MR Zbl

[15] **H Blum**, **Y Liu**, **C Xu**, **Z Zhuang**, *The existence of the Kähler–Ricci soliton degeneration*, Forum Math. Pi 11 (2023) art. id. e9 MR Zbl

[16] **H Blum**, **Y Liu**, **C Zhou**, *Optimal destabilization of K–unstable Fano varieties via stability thresholds*, Geom. Topol. 26 (2022) 2507–2564 MR Zbl

[17] **S Boucksom**, **H Chen**, *Okounkov bodies of filtered linear series*, Compos. Math. 147 (2011) 1205–1229 MR Zbl

[18] **S Boucksom**, **D Eriksson**, *Spaces of norms, determinant of cohomology and Fekete points in non-Archimedean geometry*, Adv. Math. 378 (2021) art. id. 107501 MR Zbl

[19] **S Boucksom**, **T Hisamoto**, **M Jonsson**, *Uniform K–stability, Duistermaat–Heckman measures and singularities of pairs*, Ann. Inst. Fourier (Grenoble) 67 (2017) 743–841 MR Zbl

[20] **S Boucksom**, **M Jonsson**, *A non-Archimedean approach to K–stability*, preprint (2018) arXiv 1805.11160v1

[21] **S Boucksom**, **M Jonsson**, *A non-Archimedean approach to K–stability*, I: *Metric geometry of spaces of test configurations and valuations*, preprint (2021) arXiv 2107.11221

[22] **S Boucksom**, **A Küronya**, **C Maclean**, **T Szemberg**, *Vanishing sequences and Okounkov bodies*, Math. Ann. 361 (2015) 811–834 MR Zbl

[23] **M Brion**, *Sur l'image de l'application moment*, from "Séminaire d'algèbre Paul Dubreil et Marie-Paule Malliavin" (M-P Malliavin, editor), Lecture Notes in Math. 1296, Springer (1987) 177–192 MR Zbl

[24] **H Chen**, **C Maclean**, *Distribution of logarithmic spectra of the equilibrium energy*, Manuscripta Math. 146 (2015) 365–394 MR Zbl

[25] **X Chen**, **S Donaldson**, **S Sun**, *Kähler–Einstein metrics on Fano manifolds, I: Approximation of metrics with cone singularities*, J. Amer. Math. Soc. 28 (2015) 183–197 MR Zbl

[26] **X Chen**, **S Donaldson**, **S Sun**, *Kähler–Einstein metrics on Fano manifolds, II: Limits with cone angle less than 2π*, J. Amer. Math. Soc. 28 (2015) 199–234 MR Zbl

[27] **X Chen**, **S Donaldson**, **S Sun**, *Kähler–Einstein metrics on Fano manifolds, III: Limits as cone angle approaches 2π and completion of the main proof*, J. Amer. Math. Soc. 28 (2015) 235–278 MR Zbl

[28] **X Chen**, **S Sun**, **B Wang**, *Kähler–Ricci flow, Kähler–Einstein metric, and K–stability*, Geom. Topol. 22 (2018) 3145–3173 MR Zbl

[29] **X Chen**, **B Wang**, *Space of Ricci flows, II: Part B: weak compactness of the flows*, J. Differential Geom. 116 (2020) 1–123 MR Zbl

[30] **T Darvas**, **W He**, *Geodesic rays and Kähler–Ricci trajectories on Fano manifolds*, Trans. Amer. Math. Soc. 369 (2017) 5069–5085 MR Zbl

[31] **R Dervan**, **G Székelyhidi**, *The Kähler–Ricci flow and optimal degenerations*, J. Differential Geom. 116 (2020) 187–203 MR Zbl

[32] **S K Donaldson**, *Scalar curvature and stability of toric varieties*, J. Differential Geom. 62 (2002) 289–349 MR Zbl

[33] **S K Donaldson**, *Lower bounds on the Calabi functional*, J. Differential Geom. 70 (2005) 453–472 MR Zbl

[34] **S Donaldson**, **S Sun**, *Gromov–Hausdorff limits of Kähler manifolds and algebraic geometry, II*, J. Differential Geom. 107 (2017) 327–371 MR Zbl

[35] **L Ein**, **R Lazarsfeld**, **K E Smith**, *Uniform approximation of Abhyankar valuation ideals in smooth function fields*, Amer. J. Math. 125 (2003) 409–440 MR Zbl

[36] **K Fujita**, *Optimal bounds for the volumes of Kähler–Einstein Fano manifolds*, Amer. J. Math. 140 (2018) 391–414 MR Zbl

[37] **K Fujita**, *A valuative criterion for uniform K–stability of $\mathbb{Q}$–Fano varieties*, J. Reine Angew. Math. 751 (2019) 309–338 MR Zbl

[38] **K Fujita**, **Y Odaka**, *On the K–stability of Fano varieties and anticanonical divisors*, Tohoku Math. J. 70 (2018) 511–521 MR Zbl

[39] **J Han**, **C Li**, *On the Yau–Tian–Donaldson conjecture for generalized Kähler–Ricci soliton equations*, Comm. Pure Appl. Math. 76 (2023) 1793–1867 MR Zbl

[40] **W He**, *Kähler–Ricci soliton and H–functional*, Asian J. Math. 20 (2016) 645–663 MR Zbl

[41] **T Hisamoto**, *Stability and coercivity for toric polarizations*, preprint (2016) arXiv 1610.07998

[42] **T Hisamoto**, *Mabuchi's soliton metric and relative D–stability*, preprint (2019) arXiv 1905.05948

[43] **T Hisamoto**, *Geometric flow, multiplier ideal sheaves and optimal destabilizer for a Fano manifold*, J. Geom. Anal. 33 (2023) art. id. 265 MR Zbl

[44] **M Jonsson**, **M Mustaţă**, *Valuations and asymptotic invariants for sequences of ideals*, Ann. Inst. Fourier (Grenoble) 62 (2012) 2145–2209 MR Zbl

[45] **K Kaveh**, **A G Khovanskii**, *Newton–Okounkov bodies, semigroups of integral points, graded algebras and intersection theory*, Ann. of Math. 176 (2012) 925–978 MR Zbl

[46] **K Kaveh**, **A Khovanskii**, *Convex bodies and multiplicities of ideals*, Proc. Steklov Inst. Math. 286 (2014) 268–284 MR Zbl

[47] **S J Kovács**, **Z Patakfalvi**, *Projectivity of the moduli space of stable log-varieties and subadditivity of log-Kodaira dimension*, J. Amer. Math. Soc. 30 (2017) 959–1021 MR Zbl

[48] **R Lazarsfeld**, *Positivity in algebraic geometry, II: Positivity for vector bundles, and multiplier ideals*, Ergebnisse der Math. (3) 49, Springer (2004) MR Zbl

[49] **R Lazarsfeld**, **M Mustaţă**, *Convex bodies associated to linear series*, Ann. Sci. Éc. Norm. Supér. 42 (2009) 783–835 MR Zbl

[50] **C Li**, *K-semistability is equivariant volume minimization*, Duke Math. J. 166 (2017) 3147–3218 MR Zbl

[51] **C Li**, *Minimizing normalized volumes of valuations*, Math. Z. 289 (2018) 491–513 MR Zbl

[52] **C Li**, *G–uniform stability and Kähler–Einstein metrics on Fano varieties*, Invent. Math. 227 (2022) 661–744 MR Zbl

[53] **C Li**, **Y Liu**, *Kähler–Einstein metrics and volume minimization*, Adv. Math. 341 (2019) 440–492 MR Zbl

[54] **C Li**, **Y Liu**, **C Xu**, *A guided tour to normalized volume*, from "Geometric analysis: in honor of Gang Tian's 60th birthday" (J Chen, P Lu, Z Lu, Z Zhang, editors), Progr. Math. 333, Springer (2020) 167–219 MR Zbl

[55] **C Li**, **X Wang**, **C Xu**, *Algebraicity of the metric tangent cones and equivariant K–stability*, J. Amer. Math. Soc. 34 (2021) 1175–1214 MR Zbl

[56] **C Li**, **C Xu**, *Special test configuration and K–stability of Fano varieties*, Ann. of Math. 180 (2014) 197–232 MR Zbl

[57] **C Li**, **C Xu**, *Stability of valuations: higher rational rank*, Peking Math. J. 1 (2018) 1–79 MR Zbl

[58] **C Li**, **C Xu**, *Stability of valuations and Kollár components*, J. Eur. Math. Soc. 22 (2020) 2573–2627 MR Zbl

[59] **Y Li**, **Z Li**, *Equivariant $\mathbb{R}$–test configurations and semistable limits of $\mathbb{Q}$–Fano group compactifications*, preprint (2021) arXiv 2103.06439

[60] **Y Liu**, *The volume of singular Kähler–Einstein Fano varieties*, Compos. Math. 154 (2018) 1131–1158 MR Zbl

[61] **D Mumford**, *Stability of projective varieties*, Enseign. Math. 23 (1977) 39–110 MR Zbl

[62] **A Okounkov**, *Brunn–Minkowski inequality for multiplicities*, Invent. Math. 125 (1996) 405–411 MR Zbl

[63] **W Rossmann**, *Equivariant multiplicities on complex varieties*, from "Orbites unipotentes et représentations, III" (M Andler, editor), Astérisque 173–174, Soc. Math. France, Paris (1989) 313–330 MR Zbl

[64] **G Székelyhidi**, *Filtrations and test-configurations*, Math. Ann. 362 (2015) 451–484 MR Zbl

[65] **G Székelyhidi**, *The partial C^0–estimate along the continuity method*, J. Amer. Math. Soc. 29 (2016) 537–560 MR

[66] **B Teissier**, *Valuations, deformations, and toric geometry*, from "Valuation theory and its applications, II" (F-V Kuhlmann, S Kuhlmann, M Marshall, editors), Fields Inst. Commun. 33, Amer. Math. Soc., Providence, RI (2003) 361–459 MR Zbl

[67] **G Tian**, *Kähler–Einstein metrics with positive scalar curvature*, Invent. Math. 130 (1997) 1–37 MR Zbl

[68] **G Tian**, *K–stability and Kähler–Einstein metrics*, Comm. Pure Appl. Math. 68 (2015) 1085–1156 MR Zbl

[69] **G Tian**, **S Zhang**, **Z Zhang**, **X Zhu**, *Perelman's entropy and Kähler–Ricci flow on a Fano manifold*, Trans. Amer. Math. Soc. 365 (2013) 6669–6695 MR Zbl

[70] **G Tian**, **Z Zhang**, *Regularity of Kähler–Ricci flows on Fano manifolds*, Acta Math. 216 (2016) 127–176 MR Zbl

[71] **G Tian**, **X Zhu**, *A new holomorphic invariant and uniqueness of Kähler–Ricci solitons*, Comment. Math. Helv. 77 (2002) 297–325 MR Zbl

[72] **G Tian**, **X Zhu**, *Convergence of Kähler–Ricci flow*, J. Amer. Math. Soc. 20 (2007) 675–699 MR Zbl

[73] **F Wang**, **X Zhu**, *Tian's partial C^0–estimate implies Hamilton–Tian's conjecture*, Adv. Math. 381 (2021) art. id. 107619 MR Zbl

[74] **F Wang**, **X Zhu**, *Uniformly strong convergence of Kähler–Ricci flows on a Fano manifold*, Sci. China Math. 65 (2022) 2337–2370 MR Zbl

[75] **X-J Wang**, **X Zhu**, *Kähler–Ricci solitons on toric manifolds with positive first Chern class*, Adv. Math. 188 (2004) 87–103 MR Zbl

[76] **D Witt Nyström**, *Test configurations and Okounkov bodies*, Compos. Math. 148 (2012) 1736–1756 MR Zbl

[77] **M Xia**, *On sharp lower bounds for Calabi-type functionals and destabilizing properties of gradient flows*, Anal. PDE 14 (2021) 1951–1976 MR Zbl

[78] **C Xu**, *A minimizing valuation is quasi-monomial*, Ann. of Math. 191 (2020) 1003–1030 MR Zbl

[79] **C Xu**, **Z Zhuang**, *On positivity of the CM line bundle on K–moduli spaces*, Ann. of Math. 192 (2020) 1005–1068 MR Zbl

[80] **C Xu**, **Z Zhuang**, *Uniqueness of the minimizer of the normalized volume function*, Camb. J. Math. 9 (2021) 149–176 MR Zbl

[81] **Y Yao**, *Relative Ding stability and an obstruction to the existence of Mabuchi solitons*, J. Geom. Anal. 32 (2022) art. id. 105 MR Zbl

Westlake University
Hangzhou, China

Department of Mathematics, Rutgers University
Piscataway, NJ, United States

hanjiyuan@westlake.edu.cn, chi.li@rutgers.edu

Valuations on the character variety: Newton polytopes and residual Poisson bracket

JULIEN MARCHÉ

CHRISTOPHER-LLOYD SIMON

We study the space of measured laminations ML on a closed surface from the valuative point of view. We introduce and study a notion of Newton polytope for an algebraic function on the character variety. We prove, for instance, that trace functions have unit coefficients at the extremal points of their Newton polytope. Then we provide a definition of tangent space at a valuation and show how the Goldman Poisson bracket on the character variety induces a symplectic structure on this valuative model for ML. Finally, we identify this symplectic space with previous constructions due to Thurston and Bonahon.

Introduction

The algebra of functions on the character variety   Let S be a closed oriented surface of genus $g \geq 2$. Its character variety X is the quotient of the space $\mathrm{Hom}(\pi_1(S), \mathrm{SL}_2(\mathbb{C}))$ by the equivalence relation identifying ρ_1 and ρ_2 if and only if $\mathrm{tr}\,\rho_1(\gamma) = \mathrm{tr}\,\rho_2(\gamma)$ for all $\gamma \in \pi_1(S)$. By construction, it is an affine variety whose ring of functions $\mathbb{C}[X]$ is generated by the trace functions $t_\gamma \colon \rho \mapsto \mathrm{tr}\,\rho(\gamma)$ for $\gamma \in \pi_1(S)$. The function t_γ only depends on the conjugacy class of γ up to inversion, that is, on the free homotopy class of the corresponding unoriented loop.

These trace functions are not algebraically independent: the famous identity

$$\operatorname{tr}(AB) + \operatorname{tr}(AB^{-1}) = \operatorname{tr}(A)\operatorname{tr}(B)$$

for $A, B \in \mathrm{SL}_2(\mathbb{C})$ implies, for instance, that if α and β represent simple loops intersecting once, then

$$t_\alpha t_\beta = t_\gamma + t_\delta,$$

where γ and δ are elements in $\pi_1(S)$ representing the simple curves obtained by smoothing the intersection between α and β in the two possible ways.

This phenomenon generalizes as follows. Given a multiloop α, that is, a multiset $\{\alpha_1, \ldots, \alpha_n\}$ of nontrivial loops $\alpha_i \in \pi_1(S)$, the function $t_\alpha = t_{\alpha_1} t_{\alpha_2} \cdots t_{\alpha_n}$ can be uniquely decomposed as a linear combination

$$(1) \qquad\qquad\qquad\qquad t_\alpha = \sum m_\mu t_\mu,$$

where each μ is a multicurve, that is, a (possibly empty) multiloop represented by pairwise disjoint, simple, nontrivial loops. This means that the set MC of multicurves indexes a linear basis for the algebra of characters $\mathbb{C}[X]$, which is privileged from the topological viewpoint; it is also invariant under the (algebraic) automorphism group of $\mathbb{C}[X]$, as we proved in [12].

It is an old problem to understand the algebraic structure of $\mathbb{C}[X]$, whose study was initiated by Fricke and Vogt in the late 19$^{\text{th}}$ century, and revisited in the seventies by the work of Procesi, Horowitz and Magnus among others; see Magnus [11] for a review. One approach is to investigate the coefficients m_μ of the functions t_α.

In this article, we define the Newton set $\Delta(t_\alpha) \subset \mathrm{MC}$ of t_α, in analogy with the extremal points of the ordinary Newton polytope of a polynomial, as follows.

Definition (Newton set) For $f = \sum m_\mu t_\mu$ decomposed in the basis of multicurves, we define its *support* as $\operatorname{Supp}(f) = \{\mu \in \mathrm{MC} \mid m_\mu \neq 0\}$.

We say that $\mu \in \operatorname{Supp}(f)$ is *extremal* in f if there exists a multicurve ξ such that $i(\xi, \mu) > i(\xi, \nu)$ for all $\nu \in \operatorname{Supp}(f)$ distinct from μ.

The *Newton set* $\Delta(f)$ is the set of extremal multicurves in f.

In this definition, $i(\cdot, \cdot)$ denotes the geometric intersection number, and standard properties of measured laminations imply that ξ can be replaced by a simple curve or a measured lamination. Our first result is the following.

Theorem A (trace functions are unitary) *For every multiloop $\alpha = \{\alpha_1, \ldots, \alpha_n\}$, the function t_α is unitary in the sense that $m_\mu = \pm 1$ for all $\mu \in \Delta(t_\alpha)$.*

To introduce our next result, recall that the algebra of functions $\mathbb{C}[X]$ carries a natural Poisson bracket stemming from the Atiyah–Bott–Weil–Petersson–Goldman symplectic structure on X. Following Goldman [7], for $\alpha, \beta \in \pi_1(S)$ it is given by the formula

$$\{t_\alpha, t_\beta\} = \sum_{p \in \alpha \cap \beta} \epsilon_p (t_{\alpha_p \beta_p} - t_{\alpha_p \beta_p^{-1}}), \tag{2}$$

where the sum ranges over all intersection points p between transverse representatives for $\alpha \cup \beta$, and ϵ_p is the sign of such an intersection, while α_p and β_p denote the homotopy classes of α and β based at p.

Our second result interprets the coefficients of $\{f, g\}$ at the extremal multicurves of fg in terms of Thurston's PL–symplectic structure on the space ML of measured laminations in S.

Theorem B (extremal structure constants for the Poisson bracket) *Let μ and ν be two multicurves. For $\xi \in \Delta(t_\mu t_\nu)$ we set $E_\xi = \{\lambda \in \mathrm{ML} \mid i(\xi, \lambda) = i(\mu, \lambda) + i(\nu, \lambda)\}$. These closed subsets of ML form a piecewise linear partition of ML with disjoint interiors.*

For Thurston's symplectic structure, the Poisson bracket $\{i_\mu, i_\nu\}$ of the length functions defined by $i_\mu(\lambda) = i(\mu, \lambda)$ is equal to the coefficient of t_ξ in $\{t_\mu, t_\nu\}$ almost everywhere in E_ξ.

Let us illustrate the theorem with the following example. The curves shown in Figure 1 satisfy $t_\alpha t_\beta = t_{c_1} t_{c_3} + t_{c_2} t_{c_4} - t_\gamma - t_\delta$ and $\{t_\alpha, t_\beta\} = 2t_\delta - 2t_\gamma$, so we find that $\Delta(t_\alpha t_\beta) = \{c_1 \cup c_3, c_2 \cup c_4, \gamma, \delta\}$, whereas $\Delta(\{t_\alpha, t_\beta\}) = \{\gamma, \delta\}$.

The Newton set of $t_\alpha t_\beta$ decomposes ML into 4 domains, where $i(\alpha \cup \beta, \lambda)$ is equal to the intersection of λ with $c_1 \cup c_3$ or $c_2 \cup c_4$ or γ or δ, respectively. In the interior of these domains, $\{t_\alpha, t_\beta\}$ takes the values 0, 0, −2 and 2, respectively.

Strong relations between the symplectic structures on X and ML had already been observed, for instance in Papadopoulos and Penner [17] or Sözen and Bonahon [22]. Theorem B can be related to a formula for $\{i_\mu, i_\nu\}$ obtained in Bonahon [1, Proposition 6] by degenerating Wolpert's "cosine formula". However, our approach is algebraic in the sense that it uses valuations instead of Teichmüller theory.

Beyond these two results, the purpose of this article is to investigate the space of measured laminations from the valuative viewpoint, in particular its symplectic structure. This study was motivated by a new
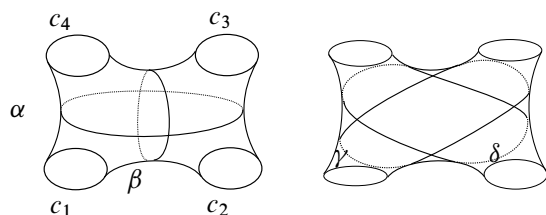

Figure 1: Product and Poisson bracket in a sphere with four punctures.

characterization of valuations associated to measured laminations that we obtained in [12]. We devote the remaining part of this introduction to an overview of our motivations, as well as the intermediate results that we obtained while revisiting the theory of measured laminations from the valuative viewpoint, since we believe they are of independent interest. We take this as an opportunity to recall general ideas for the benefit of a wide audience.

The Newton polytope   A leading analogy in this article is to think of the collection (t_μ) as a monomial basis in a polynomial algebra, keeping in mind that it is not stable under multiplication.

Consider the degree $\deg_d$ defined for $d \in \mathbb{R}^n$ on the algebra $\mathbb{C}[t_1, \ldots, t_n]$ by

$$\deg_d \left(\sum_\mu m_\mu t^\mu \right) = \max\{\langle \mu, d \rangle \mid m_\mu \neq 0\},$$

where $t^\mu = t_1^{\mu_1} \cdots t_n^{\mu_n}$, and $\langle \cdot, \cdot \rangle$ stands for the usual scalar product. This degree is (the opposite of) a monomial valuation. For $P \in \mathbb{C}[t_1, \ldots, t_n]$, a monomial t^μ is an extremal point of its usual Newton polytope $\Delta(P)$ if $m_\mu \neq 0$ and for some $d \in \mathbb{R}^n$ the maximum defining $\deg_d$ is attained uniquely at t^μ.

Our starting point is to replace the degree $\deg_d$ by the valuation associated to a measured lamination λ in S. For us a valuation will be a map $v \colon \mathbb{C}[X] \to \{-\infty\} \cup \mathbb{R}$ satisfying $v(fg) = v(f) + v(g)$ and $v(f + g) \leq \max\{v(f), v(g)\}$ for all $f, g \in \mathbb{C}[X]$. We choose this convention, which is opposite to the usual one, to avoid crowding too many signs. In the general language of valuations (see for instance Vaquié [25]), our valuations are centered at infinity on the affine variety X as they take nonnegative values on the ring $\mathbb{C}[X]$ of characters.

In a groundbreaking series of articles starting with [14], Morgan and Shalen showed that the character variety X can be compactified using valuations, in the spirit of the Riemann–Zariski compactification. In particular, the space of measured laminations, viewed as Thurston's compactification of Teichmüller space, can be embedded in the space of valuations on $\mathbb{C}[X]$ with values in an archimedean group. However, this embedding used a degeneration process and is not completely explicit: if v is the valuation associated to a lamination λ, we clearly have $v(t_\mu) = i(\lambda, \mu)$, but it was not clear what $v(f)$ should be for a general element $f \in \mathbb{C}[X]$.

In our previous article [12], we showed that the space of measured laminations ML can be identified with the space of *simple* valuations $v \colon \mathbb{C}[X] \to \{-\infty\} \cup \mathbb{R}_{\geq 0}$. The word simple means monomial with respect to the multicurve basis in the sense that the following holds:

$$(3) \qquad\qquad v\left(\sum m_\mu t_\mu \right) = \max\{v(t_\mu) \mid m_\mu \neq 0\}.$$

This justifies our definition for the Newton set of $f = \sum m_\mu t_\mu$ as the set of $\mu \in \mathrm{Supp}(f)$ such that the maximum in (3) is attained uniquely at t_μ for some $v \in \mathrm{ML}$.

For a concrete example, consider the particular case of a multiloop α contained in an incompressible pair of pants $P \subset S$. The subsurface P contains only three simple curves, its boundary components, and they

do not intersect each other. Denoting by t_1, t_2, t_3 the trace functions along these components, we have $t_\alpha \in \mathbb{Z}[t_1, t_2, t_3]$. This polynomial is often called the Fricke polynomial and has been much studied; see [11, Section 2.2]. Now any valuation associated to a measured lamination on S restricts to a monomial valuation on $\mathbb{C}[t_1, t_2, t_3]$, and we find that our Newton set corresponds to the extremal points of the usual Newton polytope. Even for such $\alpha \subset P$, it is not easy to determine $\Delta(t_\alpha)$ from the $\alpha_i \in \pi_1(S)$, and the unitarity property is not an obvious one.

It is worth noticing that we only talk about the Newton set and not about the Newton polytope, as we do not know any reasonable notion of convexity in ML. However, we can define the dual Newton polytope of a function $f \in \mathbb{C}[X]$ as $\Delta^*(f) = \{v \in \mathrm{ML} \mid v(f) \leq 1\}$. Moreover, we could define the poset of faces of $\Delta(f)$ using the order structure. Its combinatorics may be a promising land of investigation, but we did not go further in that direction.

Symplectic and combinatorial volumes of dual polytopes This paragraph only serves motivational purposes and does not claim new results; it may be skipped harmlessly.

Thurston's symplectic form on ML provides a notion of volume; thus we may ask for the topological meaning of the volume $\mathrm{Vol}\,\Delta^*(t_\alpha)$ when α is a multiloop.

When α is a filling multiloop, a celebrated theorem of M Mirzakhani [13], extended by Rafi and Souto [21], estimates the number of elements in its orbit under the modular group $\mathrm{Mod}(S)$ as a bound on their complexity tends to infinity. More precisely, fix another filling multiloop β, and denote by $m_g > 0$ the volume of the moduli space of hyperbolic metrics on S for the Weil–Petersson form. The theorem claims the following:

$$\lim_{r \to \infty} \frac{\mathrm{Card}\{\varphi \in \mathrm{Mod}(S) \mid i(\beta, \varphi(\alpha)) \leq r\}}{r^{6g-6}} = \frac{\mathrm{Vol}\,\Delta^*(t_\beta)\,\mathrm{Vol}\,\Delta^*(t_\alpha)}{m_g}.$$

The identification between measured laminations and simple valuations implies, using equation (3), that the Newton dual polytope $\Delta^*(f)$ of $f \in \mathbb{C}[X]$ equals the intersection of $\Delta^*(t_\mu)$ for $\mu \in \Delta(f)$. These "elementary cones" $\Delta^*(t_\mu) = \{v \in \mathrm{ML} \mid v(t_\mu) \leq 1\}$ are described by explicit sets of linear inequalities in any PL chart of ML, and the volume of their intersection is computable. This yields a constructive procedure to compute Mirzakhani's constant $\mathrm{Vol}\,\Delta^*(t_\alpha)$, and shows that it depends only on $\Delta(t_\alpha)$. It also shows that these volumes are rational.

A different motivation is that this Newton set, as the usual one, could have applications to the problem of counting solutions of algebraic equations in X. We wonder for instance if it helps estimating the number of solutions to a system of $6g - 6$ equations $t_{\gamma_i} = x_i$, where $\gamma_1, \ldots, \gamma_{6g-6} \in \pi_1(S)$ and $x_1, \ldots, x_{6g-6} \in \mathbb{C}$. This could have interesting applications to three-dimensional topology, for instance to evaluate the number of points in the $\mathrm{SL}_2(\mathbb{C})$–character variety of $\pi_1(M)$ for a 3–manifold M from a Heegaard decomposition.

Measured laminations as valuations In this article we study measured laminations using the tools of valuation theory. There are two well-known invariants for an archimedean valuation v: its *rational rank*,

defined as the dimension of the $\mathbb{Q}$–vector space generated by the group Λ_v of its values (that is, differences of lengths for the corresponding measured lamination), and the *transcendence degree* of its residue field k_v. These invariants are related by the celebrated *Abhyankar inequality*, $\operatorname{rat rk}(v) + \operatorname{tr deg}(k_v) \leq 6g - 6$. Here we will show the following.

Proposition A (characterizing strict valuations) *For a valuation v associated to a measured lamination λ, the following properties are equivalent:*

(i) *Distinct multicurves μ and ν have distinct lengths: $i(\lambda, \mu) \neq i(\lambda, \nu)$.*

(ii) *The residue field of $\mathbb{C}(X)$ at v has transcendence degree 0, or $k_v = \mathbb{C}$.*

(iii) *The $\mathbb{Q}$–vector space generated by the set of lengths $i(\lambda, \mu)$ for $\mu \in \mathrm{MC}$ has dimension $6g - 6$.*

The first property implies that v defines a total order on the set of multicurves, so the max in equation (3) will always be strict, which is why they deserve to be called *strict valuations*. They played a prominent role in our previous article, where we showed that almost all valuations are strict (in the measure-theoretical sense). They will be equally important in this paper, as property (ii) enables us to define the residual value at v of a function $f \in \mathbb{C}(X)$ satisfying $v(f) \leq 0$. Combined with property (iii), it shows that strict valuations are *Abhyankar* in the sense that his inequality is an equality. We wonder whether any measured lamination gives rise to an Abhyankar valuation.

We have not come across strict valuations in the literature. Instead we encounter *maximal measured laminations*, which are those whose support cannot be enlarged. In this article, we characterize the valuations associated to maximal laminations as being *acute*: for any $\alpha, \beta \in \pi_1(S) \setminus \{1\}$ we never have $v(t_\alpha t_\beta) = v(t_{\alpha\beta}) = v(t_{\alpha\beta^{-1}})$, so that these quantities are the lengths for the edges of an acute isosceles triangle. We will show that a valuation v_λ is acute if and only if any time we smooth a self-intersection of a multiloop which is taut (minimally intersecting in its homotopy class), the two resulting multiloops have distinct λ–lengths. This property plays a crucial role in the proof of the unitarity theorem. We also show that any strict valuation is acute, and wonder if the reciprocal statement is true.

Tangent spaces and Thurston's symplectic structure The space of measured laminations is a PL–manifold but does not carry any sensible smooth structure (for which intersection numbers have smooth variations), so there is no symplectic structure in the usual sense. However, Thurston showed that most points (maximal laminations) have a well-defined tangent space endowed with a nondegenerate skew-symmetric form; see [19, Chapter 3].

In this article we propose a straightforward notion for the tangent space $T_v\mathrm{ML}$ at a valuation, and show that when v is strict, it coincides with the space $\operatorname{Hom}(\Lambda_v, \mathbb{R})$, which has dimension $\operatorname{rat rk}(v) = 6g - 6$. Then we show how the Goldman Poisson bracket induces a "residual Poisson bracket" at any strict valuation v, thus endowing $T_v\mathrm{ML}$ with a symplectic structure. For future reference we shall name this

topology/geometry	dynamics	algebra
measured foliation measured geodesic laminations	action of $\pi_1(S)$ on a real tree	simple valuation
length function	translation length	trace function
filling/aperiodic lamination	free action	positive valuation
maximal lamination	trivalent tree	acute valuation
?	?	strict valuation

Table 1

model after Goldman. This uses the crucial fact that given $f, g \in \mathbb{C}[X]$, we have $v(\{f, g\}) \leq v(fg)$ for all $v \in \mathrm{ML}$. This property amounts to the inverse inclusion of the dual polytopes $\Delta^*(\{f, g\}) \supset \Delta^*(fg)$.

Finally, we provide precise identifications between this symplectic vector space and two other existing models in the literature, which we now pass under review. In the work of Morgan and Shalen, the key notion relating measured laminations and valuations is the action of $\pi_1(S)$ on real trees. We may represent this dynamical point of view as lying between the two others as in Table 1, which the reader may use as a dictionary.

For future reference, we name the symplectic vector spaces appearing naturally from each of those approaches after Thurston, Bonahon and Goldman, respectively.

Goldman's model   It is given by the residual Poisson bracket on $T_v\mathrm{ML}$, which we introduced briefly. It will be described with more detail in the body of the paper.

Thurston's model   One can associate to a maximal measured lamination λ a ramified 2–fold covering $S' \to S$, known as the orientation cover of the lamination. The group $H^1(S', \mathbb{R})$ splits into a symmetric and antisymmetric part with respect to the involution of the covering $S' \to S$. The space $H^1(S', \mathbb{R})^-$ with the cup-product form is the geometric model for $T_\lambda\mathrm{ML}$.

Bonahon's model   If we consider a trivalent real tree T with a free and minimal action of $\pi_1(S)$, we can consider the space of functions $c : V(T)^2 \to \mathbb{R}$ on the set of pairs of trivalent vertices of T which satisfy

(i) $c(x, y) = c(y, x)$,

(ii) $c(x, y) = c(x, z) + c(z, y)$ if z belongs to the geodesic joining x to y,

(iii) $c(\alpha x, \alpha y) = c(x, y)$ for all $\alpha \in \pi_1(S)$.

Again, this space has a natural antisymmetric form related to the cyclic orientation of T at every trivalent vertex. It is equivalent to the space of transverse cocycles introduced by Bonahon; see [2, page 240]. The identification between Thurston's and Bonahon's models is well known but all proofs we encountered use auxiliary structures like train tracks. At the end of the article, we provide "invariant" proofs for the following result.

Theorem C (symplectomorphisms) *There are natural isomorphisms of symplectic vector spaces between the models of Thurston, Bonahon and Goldman.*

In particular, we provide a new construction of independent interest, reminiscent of Milnor's join construction, which, starting from a trivalent real tree, produces a space homotopically equivalent to the covering S'. We may wonder which of these three symplectic identifications persist for more general actions of Fuchsian groups on real trees.

1 Background

1.1 Algebra of functions on the character variety

Let S be a closed connected and oriented surface of genus $g \geq 1$. We denote by X the character variety of S, which is the algebraic quotient of its representation variety $\mathrm{Hom}(\pi_1(S), \mathrm{SL}_2(\mathbb{C}))$ by the conjugacy action of $\mathrm{SL}_2(\mathbb{C})$, defined as the spectrum of the algebra of invariant functions:

$$\mathbb{C}[X] = \mathbb{C}[\mathrm{Hom}(\pi_1(S), \mathrm{SL}_2(\mathbb{C}))]^{\mathrm{SL}_2(\mathbb{C})}.$$

A celebrated result of Procesi presents generators and relations for this algebra (which holds for any finitely generated group). It appears in the form presented here in [3, Proposition 9.1]. For $\alpha \in \pi_1(S)$, we denote by $t_\alpha \in \mathbb{C}[X]$ the trace function given by $t_\alpha([\rho]) = \mathrm{tr}\, \rho(\alpha)$.

Theorem 1 (Procesi) *The algebra $\mathbb{C}[X]$ is generated by the t_α for $\alpha \in \pi_1(S)$. The ideal of relations is generated by $t_1 - 2$ and $t_\alpha t_\beta - t_{\alpha\beta} - t_{\alpha\beta^{-1}}$ for all $\alpha, \beta \in \pi_1(S)$.*

Definition 2 A *multiloop* in S is a class of continuous maps $f : \Gamma \to S$ from compact 1–dimensional manifolds Γ to S which are not homotopic to a constant on any component. We consider it modulo the relation declaring f equivalent to $f' : \Gamma' \to S$ when there is a homeomorphism $\phi : \Gamma \to \Gamma'$ such that $f' \circ \phi$ is homotopic to f. We allow the empty multiloop ($\Gamma = \varnothing$).

A *multicurve* is a multiloop which is represented by an embedding. We denote by MC the set of multicurves.

A multiloop amounts to a finite multiset $\{\alpha_1, \dots, \alpha_n\}$ of nontrivial conjugacy classes in $\pi_1(S)$ considered up to inversion: we define $t_\alpha = \prod_{i=1}^n t_{\alpha_i}$, in particular $t_\varnothing = 1$. The components of a multicurve must be noncontractible, simple and pairwise disjoint.

Applying the trace relation recursively to reduce the number of self intersections in multiloops, one may deduce part of the following theorem [20]. The linear independence requires more work.

Theorem 3 *The family $(t_\mu)_{\mu \in \mathrm{MC}}$ forms a linear basis of the algebra $\mathbb{C}[X]$.*

1.2 Deriving the Poisson algebra from the Kauffman algebra

The multiplication and the Poisson bracket on $\mathbb{C}[X]$ appear naturally as byproducts of the Kauffman algebra $K(S, R)$ over some ring R containing an invertible element A. Recall that a banded link in an oriented 3–manifold is the image by a tame embedding of a finite union of oriented annuli.

As an R–module, the Kauffman algebra is the quotient of the free module over isotopy classes of banded links L in $S \times [0, 1]$, by the submodule generated by Kauffman's local skein relations

$$[\bigcirc \cup L] = (-A^2 - A^{-2})[L] \quad \text{and} \quad [L_\times] = A[L_+] + A^{-1}[L_-],$$

where $L_\times, L_+, L_-$ are banded links differing in a ball as shown in Figure 2.

The product is given by stacking two banded links one above the other. Precisely,

$$[L_0][L_1] = [\Phi_0(L_0) \cup \Phi_1(L_1)], \quad \text{where } \Phi_i(x, t) = \left(x, \tfrac{1}{2}(t + i)\right).$$

Any multicurve μ on S can be seen as a banded link $[\mu]$ in $S \times [0, 1]$ by considering a tubular neighborhood $S \times \left\{\tfrac{1}{2}\right\}$, often called its blackboard framing.

We sum up what we need to know about skein algebras in the following theorem.

Theorem 4 *Using the previous notation*:

(i) *The module $K(S, R)$ is a free R–module generated by multicurves.*

(ii) *The algebra $K(S, \mathbb{C})$ with $A = -1$ is commutative, and there is an isomorphism $K(S, \mathbb{C}) \to \mathbb{C}[X]$ defined by sending the blackboard framing $[\mu]$ of a multicurve $\mu \in \mathrm{MC}$ to $(-1)^{|\mu|} t_\mu$, where $|\mu|$ denotes the number of components of μ.*

(iii) *The map sending a multicurve to its blackboard framing yields an isomorphism of $\mathbb{C}[A^{\pm 1}]$–modules $K(S, \mathbb{C}) \otimes \mathbb{C}[A^{\pm 1}] \simeq K(S, \mathbb{C}[A^{\pm 1}])$. In this setting, we have*

$$\{f, g\} = \frac{1}{2} \frac{d}{dA} \left[fg - gf\right]_{A=-1}.$$

These algebras were introduced independently by Przytycki and Turaev. The assertions in part (i), in part (ii) and the isomorphism of part (iii) are [20, Fact 4.1, Fact 2.7 and Theorem 2.8]. Part (ii) is also proved in [4]. Finally, the last formula appears in [5].
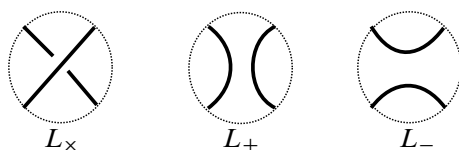

Figure 2: Local skein relation.

Let us explain Theorem 4(i) more precisely. Given a diagram D for a banded link $L \subset S \times [0, 1]$, we denote by C its set of crossings. For any map $\sigma \colon C \to \{\pm 1\}$, let $w_\sigma = \sum_c \sigma(c)$ and consider the diagram D_σ obtained after smoothing each crossing $c \in C$ according to the sign $\sigma(c)$, and removing the n_σ trivial components which appear in the result. The following formula holds in $K(S, R)$:

$$(4) \qquad [L] = \sum_{\sigma \colon C \to \{\pm 1\}} (-A^2 - A^{-2})^{n_\sigma} A^{w_\sigma} [D_\sigma],$$

which, after grouping terms corresponding to a same diagram $[D_\sigma]$, yields the decomposition of $[L]$ in the basis of multicurves. This formula sheds light on the product of two multicurves μ, ν: intuitively, the product is obtained by taking the union $\mu \cup \nu$ and summing over all possible smoothings.

By Theorem 4(ii), we deduce that the algebra $\mathbb{C}[X]$ has a linear basis indexed by trace functions of multicurves. At $A = -1$, the class of $[L]$ does not change if we change a crossing. Hence, we can replace the notion of banded link with the simpler notion of multiloop that we defined previously.

The Kauffman algebra is not completely necessary for our purposes. However, we find it conceptually useful for the following reasons. It transforms the trace relation into a local relation whose sign is more convenient (for instance while performing successive diagrammatic computations), and a better understanding of the product in terms of smoothings. It also provides a simple reason as to why the Goldman bracket actually satisfies the Jacobi relation: this comes from Theorem 4(iii) and the obvious associativity of multiplication in the Kauffman algebra. Finally, in the context of this article, it provides an alternative formula for the Poisson bracket which enlightens Theorem B. Indeed, a smoothing σ which is extremal for $[\alpha][\beta]$ in $K(S, R)$ is also extremal for the Poisson bracket $\{t_\alpha, t_\beta\}$ in $\mathbb{C}[X]$. Its coefficient in the former is ± 1 by Theorem A, and we will interpret its coefficient $\pm w_\sigma$ in the latter as a residual Poisson bracket.

2 Measured laminations and simple valuations

2.1 Simple valuations

It is well known that a measured lamination λ on S is characterized by the length $i(\lambda, \gamma)$ it assigns to every simple curve γ. This "functional" point of view can be extended to define a map $v_\lambda \colon \mathbb{C}[X] \to \{-\infty\} \cup \mathbb{R}_{\geq 0}$ satisfying $v(0) = -\infty$ and for all $f = \sum m_\mu t_\mu$ decomposed in the multicurve basis,

$$(5) \qquad v_\lambda(f) = \max\{i(\lambda, \mu) \mid m_\mu \neq 0\},$$

where $i(\lambda, \mu) = i(\lambda, \mu_1) + \cdots + i(\lambda, \mu_n)$ for a multicurve μ with components $\mu_1, \ldots, \mu_n$. By [12, Proposition 1.2], equation (5) is coherent with the fact that for any $\alpha \in \pi_1(S)$, not necessarily simple, we actually have $v_\lambda(t_\alpha) = i(\lambda, \alpha)$. Let us recall [12, Definition 1.1].

Definition 5 A simple valuation on $\mathbb{C}[X]$ is a map $v \colon \mathbb{C}[X] \to \{-\infty\} \cup \mathbb{R}_{\geq 0}$ satisfying:

   (i) $v(f) = -\infty$ if and only if $f = 0$.

(ii) $v(fg) = v(f) + v(g)$ for all $f, g \in \mathbb{C}[X]$.

(iii) If $f = \sum m_\mu t_\mu$ then $v(f) = \max\{v(t_\mu) \mid m_\mu \neq 0\}$.

The following characterization was of fundamental importance in [12]: it yields a homeomorphism between the space of simple valuations and ML, both topologies being defined by simple convergence for the evaluations of multicurves.

Theorem 6 (Marché–Simon) *The simple valuations on $\mathbb{C}[X]$ are precisely the v_λ for $\lambda \in \mathrm{ML}$.*

In this paper we only consider simple valuations, so we write $v \in \mathrm{ML}$ and $\lambda \in \mathrm{ML}$ interchangeably.

The maximality condition of Definition 5 implies that for any $f, g \in \mathbb{C}[X]$, we have $v(f + g) \leq \max\{v(f), v(g)\}$, with equality if $v(f) \neq v(g)$. Given a multiloop α with a self-intersection p, the two smoothings at p give multiloops α_+ and α_- and the trace relation reads $t_\alpha = \pm t_{\alpha_+} \pm t_{\alpha_-}$. Hence any valuation v satisfies $v(t_\alpha) \leq \max\{v(t_{\alpha_+}), v(t_{\alpha_-})\}$. The following lemma was proven by Dylan Thurston in [23], and removes the condition $v(t_{\alpha_+}) \neq v(t_{\alpha_-})$ for the equality to hold. We provide an independent proof in Section 5, which relies on the geometry of real trees.

Lemma 7 (smoothing lemma) *Let α be a taut multiloop, having a self-intersection p with smoothings α_+ and α_-. For any $v \in \mathrm{ML}$ we have $v(t_\alpha) = \max\{v(t_{\alpha_+}), v(t_{\alpha_-})\}$.*

Still, it will prove useful to consider valuations v for which we always have $v(t_{\alpha_+}) \neq v(t_{\alpha_-})$. This holds over subsets of full measure in ML, as we now explain.

2.2 Acute valuations

We say that a simple valuation $v = v_\lambda \in \mathrm{ML}$ is *positive* if $v(f) > 0$ for all nonconstant $f \in \mathbb{C}[X]$. It is equivalent to saying that $i(\lambda, \alpha) > 0$ for all $\alpha \in \pi_1(S)$, or $i(\lambda, \mu) > 0$ for all simple curves μ. Such measured laminations are called filling or aperiodic in the literature.

We now introduce the notion of acute valuation, which will happen to be equivalent to the notion of maximal measured geodesic lamination, as we will show in Proposition 27.

Definition 8 A *simple* valuation $v \in \mathrm{ML}$ is called acute if it is positive and for any nontrivial $\alpha, \beta \in \pi_1(S)$, we do not have $v(t_{\alpha\beta}) = v(t_\alpha t_\beta) = v(t_{\alpha\beta^{-1}})$.

Lemma 9 (unique smoothing) *A positive simple valuation $v_\lambda \in \mathrm{ML}$ is acute if and only if for every taut multiloop α, and smoothings $\alpha_\pm$ at a self-intersection, we have*

$$i(\lambda, \alpha_+) \neq i(\lambda, \alpha_-).$$

This justifies the terminology: $v \in \mathrm{ML}$ is acute when for every such a multiloop α, we have either $v(\alpha) = v(\alpha_+) > v(\alpha_-)$ or $v(\alpha) = v(\alpha_-) > v(\alpha_+)$, so the values $v(\alpha), v(\alpha_+), v(\alpha_-)$ are the lengths of an acute isosceles triangle with one shortest edge corresponding to either $v(\alpha_-)$ or $v(\alpha_+)$.

Proof Suppose $v \in \mathrm{ML}$ is acute. By decomposing α into connected components, we observe that the smoothing concerns at most two of them, and the proof reduces to the following cases.

(i) Either α is a single loop, self-intersecting at p. Denote by $\gamma, \delta \in \pi_1(S, p)$ the elements such that α is homotopic to $\gamma\delta$. The tautness assumption implies that γ and δ are nontrivial. Depending on the combinatorics of the intersection, one smoothing is homotopic to $\gamma\delta^{-1}$ and the other to the union $\gamma \cup \delta$. If $v(t_{\gamma\delta^{-1}}) = v(t_\gamma t_\delta)$ then, from the acute property, $v(t_{\gamma\delta})$ differs from them, which contradicts the smoothing lemma (Lemma 7).

(ii) Otherwise the multiloop α has two components intersecting at p. We denote by $\gamma, \delta \in \pi_1(S, p)$ the (nontrivial) homotopy classes of the two components. Again, α_+ and α_- are homotopic to $\gamma\delta$ and $\gamma\delta^{-1}$; the reasoning is the same.

Conversely, suppose $\alpha, \beta \in \pi_1(S)$ are nontrivial. If they are powers of a same element, say $\alpha = \gamma^n$ and $\beta = \gamma^m$, then $v(t_{\alpha\beta}) = |n + m| v(t_\gamma)$ and $v(t_{\alpha\beta^{-1}}) = |n - m| v(t_\gamma)$. As $v(t_\gamma) > 0$, the equality $v(t_{\alpha\beta}) = v(t_\alpha t_\beta) = v(t_{\alpha\beta^{-1}})$ implies $mn = 0$, which is impossible.

Consider a hyperbolic structure on S, so that α and β act on $\widetilde{S} \simeq \mathbb{H}^2$ by hyperbolic translations along distinct axes A_α and A_β, respectively.

(i) If $A_\alpha \cap A_\beta = \{p\}$, then p projects to a point on $\alpha \cap \beta$. The smoothings at p are $\alpha\beta$ and $\alpha\beta^{-1}$. The assumption $i(\lambda, \alpha\beta) \neq i(\lambda, \alpha\beta^{-1})$ says that v satisfies the condition $v(t_\alpha t_\beta) \neq v(t_{\alpha\beta^{-1}})$, ensuring that of Definition 8.

(ii) If $A_\alpha \cap A_\beta = \varnothing$, then up to replacing β with β^{-1}, we may assume the axes point in the same direction. Now, the axes of $\alpha\beta$ and $\beta\alpha$ intersect at a point p. This point projects to a self-intersection of $\alpha\beta$ which, after smoothing, gives alternatively $\alpha \cup \beta$ and $\alpha\beta^{-1}$. The assumption $i(\lambda, \alpha \cup \beta) \neq i(\lambda, \alpha\beta^{-1})$ says that v satisfies the condition $v(t_\alpha t_\beta) \neq v(t_{\alpha\beta^{-1}})$, ensuring that of Definition 8.   □

2.3 Strict valuations

A simple valuation v can be extended to $\mathbb{C}(X)$ by $v(f/g) = v(f) - v(g)$. We define its valuation ring $\mathcal{O}_v = \{f \in \mathbb{C}(X) \mid v(f) \leq 0\}$, which has a unique maximal ideal $\mathcal{M}_v = \{f \in \mathbb{C}(X) \mid v(f) < 0\}$ and residue field $k_v = \mathcal{O}_v/\mathcal{M}_v$.

Lemma 10 *A simple valuation $v = v_\lambda$ satisfies $k_v = \mathbb{C}$ if and only if for all distinct multicurves μ, ν we have $i(\lambda, \mu) \neq i(\lambda, \nu)$.*

Following [12], we will refer to them as *strict* valuations. We showed in [12, Lemma 3.4] that the set of nonstrict valuations has zero measure in ML.

Proof Suppose that $k_v = \mathbb{C}$ and consider two distinct multicurves μ and ν. If $v(t_\mu) = v(t_\nu)$ then $t_\mu/t_\nu \in \mathcal{O}_v \setminus \mathcal{M}_v$, so there exists $\lambda \in \mathbb{C}^*$ such that $t_\mu/t_\nu - \lambda \in \mathcal{M}_v$ thus $v(t_\mu/t_\nu - \lambda) < 0$. But this implies $v(t_\mu - \lambda t_\nu) < v(t_\nu)$, which contradicts the third condition in Definition 5.

Conversely, suppose that v takes distinct values on distinct multicurves and pick $f = P/Q \in \mathcal{O}_v \setminus \mathcal{M}_v$. Then $v(P) = v(Q)$, so the decompositions of P and Q in the basis of multicurves must be of the form $P = at_\mu + P'$ and $Q = bt_\mu + Q'$ with $a, b \in \mathbb{C}^*$ and $v(P'), v(Q') < v(t_\mu)$. This gives

$$f = \frac{at_\mu + P'}{bt_\mu + Q'} = \frac{a + P'/t_\mu}{b + Q'/t_\mu} = \frac{a}{b} \mod \mathcal{M}_v. \qquad \square$$

For a simple valuation $v = v_\lambda$, the set of values $\Lambda_v^+ = v(\mathbb{C}[X] \setminus \{0\})$ coincides with

$$\Lambda_v^+ = \{i(\lambda, \mu) \mid \mu \in \mathrm{MC}\}$$

by condition (iii) in Definition 5, and has the structure of an abelian semigroup by condition (ii) in Definition 5. Its associated group is $\Lambda_v = v(\mathbb{C}(X)^*)$ and consists of differences of λ–lengths.

When v is strict, the map $\mu \mapsto i(\lambda, \mu)$ is a bijection between MC and Λ_v^+. It is enlightening to think about the semigroup structure on multicurves obtained by pulling back the addition in Λ_v^+ in the following way. Let μ and ν be two multicurves, viewed as elements of $K(S, \mathbb{C})$. All smoothings of $\mu \cup \nu$ are multicurves ξ with $i(\lambda, \xi) \leq i(\lambda, \mu) + i(\lambda, \nu)$ and equality holds for exactly one of them corresponding to the "sum of μ and ν with respect to v".

We define the *rational rank* of v to be $\mathrm{rat\,rk}(v) = \dim_\mathbb{Q} \Lambda_v \otimes \mathbb{Q}$. It satisfies the following Abhyankar inequality (see [16])

$$\mathrm{rat\,rk}(v) + \mathrm{tr\,deg}(k_v) \leq \dim X,$$

from which we deduce that if a simple valuation has maximal rational rank, that is $\mathrm{rat\,rk}(v) = \dim X$, then it is strict.

Proof of Proposition A By Lemma 10, we know that the first two properties of the proposition are equivalent. The Abhyankar inequality gives the implication $\mathrm{rat\,rk}(v) = 6g - 6 \Longrightarrow \mathrm{tr\,deg}(k_v) = 0$. The reverse implication will follow from the results of the remaining sections. Precisely, given a strict valuation v, we will define a tangent space $T_v\mathrm{ML}$ whose dimension is $\mathrm{rat\,rk}(v)$. Then, we will show successively that this tangent space is isomorphic to the Bonahon and Thurston models. It is well known that the latter has dimension $6g - 6$, proving the last step of the proposition. $\qquad \square$

3 Newton polytopes of trace functions

This section relies on the following lemma, whose proof is postponed to Section 5.

Lemma 11 *The set of acute valuations has full measure in* ML.

Definition 12 Let $v \in \mathrm{ML}$ be any simple valuation and $f \in \mathbb{C}[X]$ any function decomposed as $\sum m_\mu t_\mu$ in the multicurve basis.

- The multicurve $\mu \in \mathrm{Supp}(f)$ is v–*extremal* in f if $v(t_v) < v(t_\mu)$ for every other $v \in \mathrm{Supp}(f)$.

- The multicurve μ is *extremal* in f if it is v–extremal in f for some v.

- The *Newton set* of f is the subset $\Delta(f) \subset \mathrm{MC}$ of extremal curves in f.

- The function f is *unitary* if $m_\mu = \pm 1$ for any extremal multicurve in f.

Observe that if v is strict, then μ is v–extremal in f if and only if $v(f) = v(t_\mu)$. Moreover, the density of strict valuations in ML implies that a multicurve is extremal in f if and only if it is v–extremal in f for some strict v.

3.1 Trace functions are unitary

Theorem 13 (unitarity) *If α is a multiloop in S, then t_α is unitary.*

Proof Let v be a strict acute valuation and μ be the unique multicurve such that $v(t_\alpha) = v(t_\mu)$. We must prove that $m_\mu = \pm 1$. We proceed by induction on the number of intersections of α. If there are none, then the result is obvious. Otherwise, put α in taut position and consider its smoothings at an intersection. Lemma 7 and the assumption that v is acute imply that $v(t_{\alpha_+}) \neq v(t_{\alpha_-})$. One can suppose that $v(t_\alpha) = v(t_{\alpha_+}) > v(t_{\alpha_-})$. The coefficient of t_μ in t_α is the same as in $\pm t_{\alpha_+}$, so the induction hypothesis yields the result. $\square$

Remark If we represent a taut multiloop as the projection of a banded link L in $S \times [0, 1]$, we may decompose it in the basis of multicurves $\mu \in K(S, \mathbb{Z}[A^{\pm}])$ with blackboard framing. Then, the coefficient of μ in L is equal to $A^{n^+ - n^-}$, where $n^\pm$ counts the number of $\pm$–resolutions performed while transforming L into μ. At $A = -1$, we find the sign $(-1)^s$ for the extremal coefficient, where s is the number of self-intersections of α. The proof is the same, using the skein relation inductively.

Remark We know from [24] that MC indexes another basis (t'_μ) of $\mathbb{C}[X]$ for which the multiplicative structure constants are positive. The change of basis from (t_μ) to (t'_μ) is triangular, in the sense that if $\mu = \{\mu_1, \ldots, \mu_k\}$ as a multiset, then t'_μ is a polynomial in the t_{μ_j} with leading monomial $\pm t_{\mu_1} \cdots t_{\mu_k}$. In this basis, the analogous notion of Newton set will be the same (that is, indexed by the same multicurves), and its extremal coefficients will be 1.

Corollary 14 *Any strict valuation is acute.*

Proof Let v be a strict valuation and consider a taut multiloop α. Suppose $v(t_{\alpha_+}) = v(t_{\alpha_-})$. Then t_{α_+} and t_{α_-} must have the same v–extremal multicurve μ. This defines an open condition on $v \in \mathrm{ML}$, namely that $v(t_\mu) > v(t_v)$ for all $v \in \Delta(t_{\alpha_-} t_{\alpha_+}) \setminus \{\mu\}$. But simple acute valuations are dense in ML so the same will hold for some acute valuation, contradicting Lemma 9. The conclusion follows from the converse part of that lemma. $\square$

3.2 Extremal multicurves of $t_\mu t_\nu$ and $\{t_\mu, t_\nu\}$

Let μ and ν be multicurves in S and consider a taut immersion $\mu \cup \nu$ for their union. Note that such an immersion is unique up to isotopy and permutations of parallel strands. This follows from the methods and results of [8], specifically Theorem 2.1 and the discussion following Example 2.4.

We define the embedding $L_\mu(\nu)$ obtained by smoothing all intersections of $\mu \cup \nu$ with a left turn as we travel along a segment of μ and meet a segment of ν. Smoothing all intersections with a right turn would yield $L_\nu(\mu)$.

This is the product considered by Luo in [10]; in particular his Lemma 8.1 shows that $L_\mu(\nu)$ is a multicurve (it has no trivial components) and his Theorem 2.1 describes several of its properties.

Proposition 15 *Let μ and ν be multicurves. The multicurves $L_\mu(\nu)$ and $L_\nu(\mu)$ are extremal for the product $t_\mu t_\nu$, and if $i(\mu, \nu) > 0$ then they are distinct.*

Proof If $i(\mu, \nu) = 0$ then $L_\mu(\nu) = \mu \cup \nu = L_\nu(\mu)$ and the statement follows.

Now suppose $i(\mu, \nu) > 0$. We first observe that among all smoothings of the union $\mu \cup \nu$, those which maximize v_μ are precisely $L_\mu(\nu)$ and $L_\nu(\mu)$. Indeed, we know from [10, Theorem 2.1(iii)] that $i(\mu, L_\mu(\nu)) = i(\mu, \nu) = i(\mu, L_\nu(\mu))$, but any other smoothing ξ is made of segments of μ and ν which somewhere alternate between a left turn and right turn, thus forming a bigon with μ so that $i(\mu, \xi) < i(\mu, \nu)$. The fact that $L_\mu(\nu) \neq L_\nu(\mu)$ can be obtained from [10, Corollary 8.2], which proves $i(L_\mu(\nu), L_\nu(\mu)) = 2i(\mu, \nu)$.

We deduce from the preceding discussion and the multiplication formula (4) that the distinct multicurves $L_\mu(\nu)$ and $L_\nu(\mu)$ both appear in the decomposition of $t_\mu t_\nu$, and are the only two maximizers of v_μ. The condition $v_\lambda(L_\mu(\nu)) = v_\lambda(L_\nu(\mu))$ defines on $\lambda \in \mathrm{ML}$ a codimension-1 PL–subset; see [12, Lemma 1.6] for a proof. Hence a slight perturbation of the valuation v_μ off that subset in one direction or the other shows that $L_\mu(\nu)$ and $L_\nu(\mu)$ are indeed extremal terms in the product. $\square$

Corollary 16 *If μ and ν are multicurves such that $i(\mu, \nu) > 0$, then $L_\mu(\nu)$ and $L_\nu(\mu)$ are extremal in the Poisson bracket $\{t_\mu, t_\nu\}$, and their coefficients in the basis of multicurves are equal to $\pm i(\mu, \nu)$.*

Proof We may deduce this using Theorem 4, which derives the Poisson bracket from the commutator in the skein algebra, but let us detail the computation without referring to the skein product.

For this, apply the Goldman formula (2) to the multiloops α and β, and for each $p \in \alpha \cap \beta$, decompose the terms $t_{\alpha_p \beta_p}$ and $t_{\alpha_p \beta_p^{-1}}$ in the basis of multicurves $\xi \in \mathrm{MC}$, to find

$$\{t_\alpha, t_\beta\} = \sum_\xi w_\xi t_\xi = \sum_\xi \left(\sum_{\sigma_\xi} \prod_p \sigma_\xi(p) \right) t_\xi,$$

where $w_\xi = \sum_{\sigma_\xi} \prod_p \sigma_\xi(p)$ is the sum over the smoothings $\sigma_\xi \colon \alpha \cap \beta \to \{\pm 1\}$ of $\alpha \cup \beta$ yielding the multiloop ξ.

Now suppose that $\alpha = \mu$ and $\beta = \nu$ are multicurves with $i(\mu, \nu) > 0$, and consider the multicurves ξ indexing the sum that are obtained by smoothing all intersections of $\mu \cup \nu$. Reasoning as in the proof of Proposition 15, we find that $L_\mu(\nu)$ and $L_\nu(\mu)$ both index a term corresponding to a unique smoothing map σ_ξ which is constant, equal to 1 or −1.   □

Remark In the next section, we will prove that extremal coefficients of $\{t_\mu, t_\nu\}$ which are also extremal for $t_\mu t_\nu$ are values of the Thurston Poisson bracket $\{i_\mu, i_\nu\}_\lambda$ for $\lambda \in \mathrm{ML}$, as announced in Theorem B. Our approach consists in reinterpreting the Thurston Poisson bracket $\{i_\mu, i_\nu\}_\lambda$ as a residual value $\{t_\mu, t_\nu\}_v$ of the Goldman Poisson bracket at $v = v_\lambda$.

The previous corollary shows that the (residual) Poisson bracket of multicurves determines their intersection number by the formula
$$i(\mu, \nu) = \max\{\{i_\mu, i_\nu\}_\lambda \mid \lambda \in \mathrm{ML}\}.$$

4 Residual Poisson structure on ML

4.1 Tangent space

Recall that ML embeds in the space of real functions on $\mathbb{C}[X]^* = \mathbb{C}[X] \setminus \{0\}$. We thus define its tangent space at v as the set of maps
$$\phi = \left.\frac{d}{ds}\right|_{s=0} v_s \colon \mathbb{C}[X]^* \to \mathbb{R},$$
where v_s is a family of simple valuations depending on a parameter $s \in [0, \epsilon[$ starting at $v_0 = v$, such that the map $s \mapsto v_s(t_\gamma)$ is differentiable for every curve γ.

Observe that the pair $(v, \phi) \colon \mathbb{C}[X]^* \to [0, +\infty) \times \mathbb{R}$ satisfies all the axioms in Definition 5 of simple valuations provided the maximum is taken with respect to the lexicographic ordering. When v is a strict valuation, the lexicographic ordering depends only on the first coordinate and everything becomes much easier. As we only deal with the strict case, we consider straight away the following as a definition.

Definition 17 Let $v \in \mathrm{ML}$ be a strict valuation. We define $T_v\mathrm{ML}$ to be the set of group homomorphisms $\phi \colon \mathbb{C}(X)^* \to \mathbb{R}$ satisfying the property that, for any function $f \in \mathbb{C}[X]$ decomposed as $f = \sum m_\mu t_\mu$ in the linear basis of multicurves,

(6) $\phi(f) = \phi(t_\nu)$,   where ν is v–extremal in f.

We will refer to this definition of the tangent space as the Goldman model. In this section, we define a symplectic structure on it, and will relate it to the models of Thurston and Bonahon introduced later on.

Proposition 18 *For any strict valuation we have a sequence of natural isomorphisms*

$$T_v\mathrm{ML} = \mathrm{Hom}(\Lambda_v^+, \mathbb{R}) = \mathrm{Hom}(\Lambda_v, \mathbb{R}) = \mathrm{Hom}(\Lambda_v \otimes \mathbb{Q}, \mathbb{R}) = \mathrm{Hom}(\mathbb{C}(X)^*/\mathcal{O}_v^\times, \mathbb{R}),$$

where Hom *is understood first as the space of semigroup homomorphisms, and then as the space of group homomorphisms. In particular, $T_v\mathrm{ML}$ has dimension* $\mathrm{rat}\,\mathrm{rk}(v)$ *(which is $\leq \dim X$).*

Proof Recall that the map $\mu \mapsto v(t_\mu)$ is a bijection between the set of multicurves and Λ_v^+. We have $v(t_\mu) + v(t_\nu) = v(t_\xi)$, where ξ is the v–extremal multicurve in $t_\mu t_\nu$. Given $\phi \in T_v\mathrm{ML}$, the map $v(t_\mu) \mapsto \phi(t_\mu)$ is by construction a homomorphism of semigroups $\Lambda_v^+ \to \mathbb{R}$, and this construction can easily be reversed, giving the isomorphism $T_v\mathrm{ML} = \mathrm{Hom}(\Lambda_v^+, \mathbb{R})$. The remaining isomorphisms are purely formal, noticing that $\mathcal{O}_v^\times$ is the kernel of the group homomorphism $v\colon \mathbb{C}(X)^* \to \mathbb{R}$. □

Definition 19 For $f \in \mathbb{C}(X)$, we define the *differential* of the map $v \mapsto v(f)$ at v by

$$d_v \log f : T_v\mathrm{ML} \to \mathbb{R}, \quad d_v \log f(\phi) = \phi(f).$$

We introduced the log to make the formula $d_v \log(fg) = d_v \log f + d_v \log g$ look more natural.

By Proposition 18, the elements $d_v \log f$ span $T_v^*\mathrm{ML}$. More precisely, we obtain a basis by letting f range over a family of multicurves whose v–lengths form a basis of $\Lambda_v \otimes \mathbb{Q}$.

4.2 Residual Poisson structure

Proposition 20 *For all $f, g \in \mathbb{C}[X]$ and $v \in \mathrm{ML}$, we have $v(\{f, g\}) \leq v(f) + v(g)$.*

Proof By linearity of the Poisson bracket, it is sufficient to prove the inequality for $f = t_\mu$ and $g = t_\nu$, where μ and ν are multicurves. Then, by the Leibnitz formula, it is sufficient to prove it for curves μ and ν. Suppose that μ and ν are in taut position and apply Goldman's formula (2). It is sufficient to prove that for any $p \in \mu \cap \nu$ we have $v(t_{\mu_p \nu_p} - t_{\mu_p \nu_p^{-1}}) \leq v(t_\mu t_\nu)$, but this is a consequence of the smoothing lemma (Lemma 7). □

Given a strict valuation $v \in \mathrm{ML}$, the preceding proposition allows us to define the residual Poisson bracket at v in the following way.

Definition 21 For $f, g \in \mathbb{C}[X]$ and $v \in \mathrm{ML}$ strict, we define $\{f, g\}_v \in k_v = \mathbb{C}$ by

$$\{f, g\}_v = \frac{\{f, g\}}{fg} \mod \mathcal{M}_v.$$

Proposition 22 *There is an element $\pi_v \in \Lambda^2 T_v\mathrm{ML}$ representing this Poisson structure, in the sense that for any $f, g \in \mathbb{C}[X]$, we have*

$$\{f, g\}_v = \langle \pi_v, d_v \log(f) \wedge d_v \log(g) \rangle.$$

Proof Let us fix f and consider the map $F\colon \mathbb{C}[X] \to \mathbb{C}$ defined by $F(g) = \{f, g\}_v$. By the Leibnitz identity, this map satisfies $F(g_1 g_2) = F(g_1) + F(g_2)$ and thus extends to an element of $\mathrm{Hom}(\mathbb{C}(X)^*, \mathbb{C})$, and we must first show that it vanishes on $\mathcal{O}_v^\times$. Any $g \in \mathcal{O}_v^\times$ can be written $g = \alpha + h$ with $\alpha \in \mathbb{C}^*$ and $v(h) < 0$. We compute

$$F(g) = \frac{\{f, \alpha + h\}}{f(\alpha + h)} = \frac{\{f, h\}}{f(\alpha + h)}.$$

Since $v(h) < 0$ we have $v(f(\alpha+h)) = v(f) + v(\alpha+h) = v(f)$, and with Proposition 20, $v(\{f, h\}) < v(f)$; thus $F(g) \in \mathcal{M}_v$, and the claim is proved. What we have shown implies that there exists an element $\phi_f \in T_v\mathrm{ML}$ such that $F(g) = \langle \phi_f, d_v \log g \rangle$. As the Poisson bracket is antisymmetric, the same is true with the variables interchanged, and the conclusion follows.   □

5 Actions of $\pi_1(S)$ on real trees

A real tree is a metric space T such that any two points $x, y \in T$ are joined by a unique injective segment. Recall that S is a closed oriented surface of genus $g \geq 2$. We consider real trees with an action of $\pi_1(S)$ that is *minimal*, in the sense that the only subtrees $T' \subset T$ satisfying $\gamma T' \subset T'$ for all $\gamma \in \pi_1(S)$ are $\varnothing$ and T.

The action of an element $\alpha \in \pi_1(S)$ on T either fixes a point and is called elliptic; otherwise it is a hyperbolic translation along an axis A_α with positive translation length $l(\alpha) = \min\{d(x, \alpha x) \mid x \in T\}$, and $d(x, \alpha x) = l(\alpha)$ if and only if $x \in A_\alpha$.

We face the following alternative. If all elements of $\pi_1(S)$ act elliptically, then they have a common fixed point; the minimality assumption implies that T is reduced to a point. If at least one element of $\pi_1(S)$ acts hyperbolically, then the union of all translation axes forms an invariant subtree (see [18]), which equals T by the minimality assumption.

An action of $\pi_1(S)$ is *free* when only the trivial element of $\pi_1(S)$ has a fixed point, or equivalently when $l(\alpha) > 0$ for all nontrivial $\alpha \in \pi_1(S)$. It is *small* when the stabilizer of any nontrivial segment in T is cyclic. This condition appears naturally in the following important results.

Theorem 23 (Culler and Morgan) *For real trees T_1 and T_2 with small minimal actions of $\pi_1(S)$, there exists an equivariant isometry $\Phi\colon T_1 \to T_2$ if and only if $l_1(\alpha) = l_2(\alpha)$ for all $\alpha \in \pi_1(S)$.*

Theorem 24 (Thurston, Skora) *To any measured lamination $\lambda \in \mathrm{ML}$ one can associate a "dual tree" T_λ together with a small minimal action of $\pi_1(S)$ on T_λ such that $l(\alpha) = 2i(\lambda, \alpha)$ for all $\alpha \in \pi_1(S)$. Conversely, any tree with a small and minimal action of $\pi_1(S)$ is produced in this way.*

Let us briefly outline the construction of the dual tree to a measured lamination, in the case where λ is filling (or equivalently when the simple valuation v_λ is positive).

First represent the filling measured lamination λ on S by a measured geodesic lamination for some fixed hyperbolic metric, and lift it in $\widetilde{S}$ to obtain a $\pi_1(S)$–invariant measured geodesic lamination $\widetilde{\lambda}$. Following [15, Section 2.3], the tree T_λ is the quotient of $\widetilde{S}$ by the equivalence relation whose classes are given either by the closure of a connected component of $\widetilde{S} \setminus \widetilde{\lambda}$ or else by a leaf of $\widetilde{\lambda}$ which is not contained in the previous classes. The quotient map $f \colon \widetilde{S} \to T_\lambda$ is clearly $\pi_1(S)$–equivariant.

To describe the complement of a point $x \in T_\lambda$, consider its preimage $f^{-1}(\{x\})$. If it consists of a geodesic leaf of $\widetilde{\lambda}$, then $T_\lambda \setminus \{x\}$ has two connected components. Otherwise it is isometric to the closure of an ideal hyperbolic polygon with k sides, so $T_\lambda \setminus \{x\}$ has $k > 2$ components, and x is called a *branch point* of T. In any case the connected components of $T_\lambda \setminus \{x\}$ have a cyclic orientation which is $\pi_1(S)$–invariant. These local cyclic orientations match together to give a global cyclic orientation on the Gromov boundary of T_λ. See [26] for more details.

The map $f \colon \widetilde{S} \to T_\lambda$ is not proper, so does not extend to the Gromov boundary. A nontrivial element $\alpha \in \pi_1(S)$ acts on $\widetilde{S} \simeq \mathbb{H}^2$ by hyperbolic translation along an axis which is transverse to λ, and thus crosses every leaf at most once. Hence the projection f maps it bijectively to a geodesic in T which, by equivariance, coincides with the axis A_α. Hence we can associate to the attractive and repulsive points of α in $\partial \mathbb{H}^2 = \partial \pi_1(S)$ the corresponding endpoints of A_α in ∂T. This partially defined map between the Gromov boundaries of $\pi_1(S)$ and T is $\pi_1(S)$–equivariant, orientation-preserving and independent of the initial hyperbolic metric.

We recall the following proposition from [6], which we will use repeatedly.

Proposition 25 *Let γ and δ be two hyperbolic isometries acting on a real tree T with axes A_γ and A_δ. Then one of the following holds.*

(i) *If $A_\gamma \cap A_\delta = \varnothing$ then $l(\gamma\delta) = l(\gamma) + l(\delta) + 2D$ where D is the distance between A_γ and A_δ.*

(ii) *If $A_\gamma \cap A_\delta \neq \varnothing$, we denote by $D \in [0, +\infty]$ the length of the intersection.*

  (a) *If $D > 0$ and the translation directions of γ and δ on $A_\gamma \cap A_\delta$ coincide, or if $D = 0$, then $l(\gamma\delta) = l(\gamma) + l(\delta)$.*

  (b) *If $D > 0$ and the translation directions of γ and δ on $A_\gamma \cap A_\delta$ are opposite, then we have $l(\gamma\delta) < l(\gamma) + l(\delta)$.*

Corollary 26 *Let γ and δ be two hyperbolic isometries acting on a real tree T with axes A_γ and A_δ. When the segment $A_\gamma \cap A_\delta$ has positive length, we may compare the translation directions of γ and δ: let $\mathrm{cosign}(\gamma, \delta) = \pm 1$ be $+1$ if they coincide and -1 if they differ. One of the following holds:*

$(\gamma \cup \delta)$	$l(\gamma) + l(\delta) < l(\gamma\delta) = l(\gamma\delta^{-1})$	if $A_\gamma \cap A_\delta = \varnothing$.
(equil)	$l(\gamma\delta) = l(\gamma) + l(\delta) = l(\gamma\delta^{-1})$	if $A_\gamma \cap A_\delta$ is reduced to a point.
$(\gamma\delta^{-1})$	$l(\gamma\delta^{-1}) < l(\gamma) + l(\delta) = l(\gamma\delta)$	if $l(A_\gamma \cap A_\delta) > 0$ and $\mathrm{cosign}(\gamma, \delta) = 1$.
$(\gamma\delta)$	$l(\gamma\delta) < l(\gamma) + l(\delta) = l(\gamma\delta^{-1})$	if $l(A_\gamma \cap A_\delta) > 0$ and $\mathrm{cosign}(\gamma, \delta) = -1$.

To illustrate how we will apply this corollary, let us propose a new proof of the smoothing lemma, which does not rely on the equivalence between measured laminations and simple valuations (that we showed in [12] using the smoothing lemma).

Notice that the equivalence between measured laminations λ and simple valuations v recovers the smoothing lemma, because for a taut multiloop α we have for obvious geometric reasons $i(\lambda, \alpha) \geq \max\{i(\lambda, \alpha_+), i(\lambda, \alpha_-)\}$, and as $t_\alpha = \pm t_{\alpha_+} \pm t_{\alpha_-}$ we have $v_\lambda(t_\alpha) \leq \max\{v_\lambda(t_{\alpha_+}), v_\lambda(t_{\alpha_-})\}$.

Proof of the smoothing lemma (Lemma 7)   Let us represent our measured lamination λ by an action of $\pi_1(S)$ on a tree T. Fix a hyperbolic metric on S to identify $\widetilde{S} \simeq \mathbb{H}^2$.

Consider a taut multiloop α with a self-intersection point p, which may either be a self-intersection of a single component or an intersection between two components. We wish to prove that $i(\lambda, \alpha) = \max\{i(\lambda, \alpha_+), i(\lambda, \alpha_-)\}$.

Suppose first that p is the intersection point between two components which we write as $\gamma, \delta \in \pi_1(S, p)$. Lift γ and δ in $\widetilde{S} \simeq \mathbb{H}^2$ starting from $\widetilde{p}$ to obtain geodesics $\widetilde{\gamma}$ and $\widetilde{\delta}$ which intersect only at $\widetilde{p}$ and transversely at $\widetilde{p}$. Consequently, their endpoints are linked in $\partial \mathbb{H}^2$ with respect to the cyclic orientation. As the same holds for the endpoints of A_γ and A_δ in ∂T, we must have $A_\gamma \cap A_\delta \neq \varnothing$, so we are not in case ($\gamma \cup \delta$) of Corollary 26, whence $l(\gamma) + l(\delta) = \max\{l(\gamma\delta), l(\gamma\delta^{-1})\}$.

Suppose now that p is the self-intersection point of a single component of α which we may decompose as $\gamma\delta$ for $\gamma, \delta \in \pi_1(S, p)$. Lift γ and δ in $\mathbb{H}^2$ starting from $\widetilde{p}$ to obtain quasigeodesics $\widetilde{\gamma}$ and $\widetilde{\delta}$. They intersect only at $\widetilde{p}$ because, using the monodromy homomorphism associated to the developing map, another intersection point would imply an equality of the form $\gamma^m = \delta^n$ for some $m, n > 0$, which is impossible. The projection map $\widetilde{S} \to S$ is a local diffeomorphism, so the germs of arcs $(\widetilde{\gamma} \cup \widetilde{\delta}, \widetilde{p})$ and $(\gamma \cup \delta, p)$ are topologically equivalent. Hence, up to inversion and exchange of γ and δ, the endpoints of $\widetilde{\gamma}$ and $\widetilde{\delta}$ in $\partial \mathbb{H}^2$ have cyclic order $(\gamma_+, \gamma_-, \delta_+, \delta_-)$ as shown in Figure 3. Consequently we are not in case ($\gamma\delta$) of Corollary 26, whence $l(\gamma\delta) = \max\{l(\gamma) + l(\delta), l(\gamma\delta^{-1})\}$. $\square$
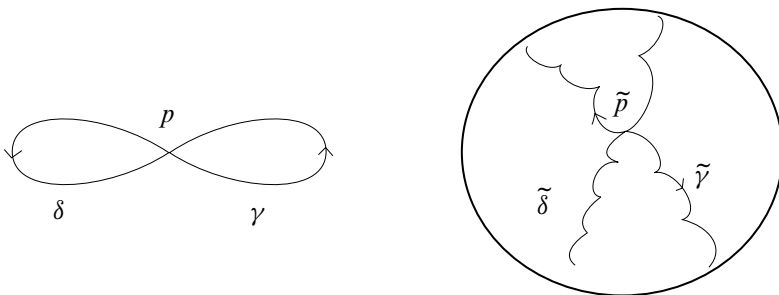

Figure 3: Configuration of axes at a self-crossing.

5.1 Trivalent real trees

Recall that a point x in a real tree is a *branch point* if $T \setminus \{x\}$ has at least three connected components. We will denote by $V(T)$ the set of branch points of T. A real tree is *trivalent* if any branch point disconnects it into three connected components.

A measured geodesic lamination λ is called maximal if there is no measured geodesic lamination whose support is strictly bigger; or equivalently if the regions in its complement $S \setminus \lambda$ are isometric to the interiors of ideal hyperbolic triangles.

Proposition 27 *Let T be a real tree with a free minimal action of $\pi_1(S)$, associated to a filling measured lamination λ. Denote by v the associated positive valuation.*

The following are equivalent:

   (i)   *v is acute.*

  (ii)   *T is trivalent.*

 (iii)   *λ is maximal.*

Proof (1) $\Longleftrightarrow$ (2) Suppose T is trivalent. Let α and β be nontrivial elements in $\pi_1(S)$ and consider their translation axes $A_\alpha, A_\beta \subset T$. From Corollary 26, we find that $l(\alpha\beta) = l(\alpha\beta^{-1}) = l(\alpha) + l(\beta)$ holds only when A_α and A_β meet in exactly one point, which is forbidden by the trivalence assumption. Thus v is acute.

Conversely, suppose T is not trivalent. Consider a branch point $x \in T$ with valency $k > 3$. We denote by $C_1, \ldots, C_k$ the components of $T \setminus \{x\}$. They decompose the Gromov boundary of T into disjoint open subsets $\partial C_1, \ldots, \partial C_k$. It is known that the set of pairs of ends of axes A_γ for $\gamma \in \pi_1(S)$ form a dense subset of $\partial T \times \partial T$. One proof consists in considering the sequence of fixed points for the elements $\alpha\beta^n$: the attractive points converge to the image by α of the attractive point of β and the repulsive points to the repulsive point of β. By minimality, the set of repulsive points of all β's is dense in ∂T, and again by minimality the images of a given attractive point by all α's is dense in ∂T. Thus we can find two axes A_α and A_β whose ends are respectively in $\partial C_1 \times \partial C_3$ and $\partial C_2 \times \partial C_4$. These two axes meet exactly at x, and Corollary 26 implies that $l(\alpha\beta) = l(\alpha\beta^{-1}) = l(\alpha) + l(\beta)$, showing that v is not acute.

(3) $\Longleftrightarrow$ (2) Recall from the construction of the dual tree T to a filling measured geodesic lamination $\lambda \subset S$ that the valency of a branch point in T is equal to the number of sides of the corresponding hyperbolic ideal polygon in $\widetilde{S} \setminus \widetilde{\lambda}$. Hence T is trivalent if and only if λ is maximal. $\qquad\square$

It is well known that the set of maximal laminations has a full measure in ML; see [9, Lemma 2.3]. Hence Proposition 27 implies the following corollary.

Corollary 28 *The set of acute simple valuations has a full measure in* ML.

5.2   Bonahon cycles

Let T be a trivalent real tree with a free and minimal action of $\pi = \pi_1(S)$. To define its tangent space in the "moduli space" of such objects, imagine the combinatorial structure as being fixed while the distance function undergoes an infinitesimal deformation. Restricting attention to the variation of the distance between branch points, we obtain a symmetric map $c \colon V(T)^2 \to \mathbb{R}$, which is $\pi_1(S)$–invariant and satisfies $c(x, y) = c(x, z) + c(z, y)$ whenever z belongs to the geodesic joining x to y. We will refer to these maps as Bonahon cocycles and introduce them formally using a dual approach.

Definition 29 We define the space $\mathcal{B}(T)$ as the real vector space generated by pairs (x, y) of elements in $V(T)$ subject to the following relations:

  (i) $(x, y) = (y, x)$ for all $x, y \in V(T)$.

  (ii) $(x, y) = (x, z) + (z, y)$ if z belongs to the geodesic joining x to y.

The group $\pi = \pi_1(S)$ acts linearly on $\mathcal{B}(T)$ by $g(x, y) = (gx, gy)$, and Bonahon cocycles are the elements of $\mathrm{Hom}_\pi(\mathcal{B}(T), \mathbb{R}) = \mathrm{Hom}(\mathcal{B}(T)_\pi, \mathbb{R})$, where $\mathcal{B}(T)_\pi$ is the space of coinvariants.

Proposition 30 *There is a unique alternating bilinear form $\cdot$ on $\mathcal{B}(T)$ such that for all pairs (x, y) and (z, t) in $V(T)^2$ we have:*

  (i) $(x, y) \cdot (z, t) = 0$ *if the geodesics from x to y and from z to t are disjoint.*

  (ii) $(x, z) \cdot (z, y) = \frac{1}{2}\epsilon$ *if z belongs to the geodesic from x to y, where $\epsilon = \pm 1$ is the cyclic order of the components (h_x, h, h_y) of $T \setminus \{z\}$ such that $x \in h_x$ and $y \in h_y$.*

Proof The intersection of (x, y) and (z, t) is either empty or has the form (a, b). Decomposing (x, y) and (z, t) into segments involving a and b as in Figure 4, we are reduced, by bilinearity and antisymmetry, to cases (i) or (ii). This proves both uniqueness and existence. $\square$

It is an amusing exercise to show that this pairing is nondegenerate. Instead we will deduce it from Poincaré duality in Thurston's model in Section 6. Indeed, we are interested in the space $\mathcal{B}(T)_\pi$ endowed with the following pairing obtained by averaging the previous one, whose nondegeneracy will thus follow from standard arguments in cohomology.
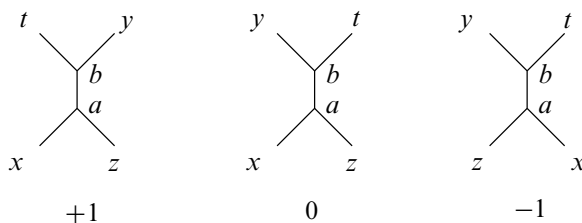

Figure 4

Proposition 31 *The following sum is finite, and it defines an alternating bilinear pairing on $\mathcal{B}(T)_\pi$:*

$$(x, y) \cdot_\pi (z, t) = \sum_{g \in \pi_1(S)} (x, y) \cdot g(z, t).$$

Proof We only have to prove finiteness of the sum. For that, we view T as the dual tree to a maximal measured lamination λ on S. The vertices x, y, z, t correspond to ideal triangles in $\widetilde{S} \simeq \mathbb{H}^2$: choose x_0, y_0, z_0, t_0 in each one of them. Since $\pi_1(S)$ acts properly on $\mathbb{H}^2$, the geodesics $[x_0, y_0]$ and $g[z_0, t_0]$ are disjoint for all but a finite number of $g \in \pi_1(S)$. When they are disjoint, their projections in the tree are disjoint or meet as in the middle case of Figure 4, so their intersection vanishes. $\qquad\square$

We shall prove in Section 6 that $\mathcal{B}(T)_\pi$ is the antisymmetric part of $H_1(\widetilde{S}, \mathbb{R})$, where $\widetilde{S}$ is the orientation covering of the measured lamination λ, thus recovering Thurston's original point of view on the tangent space $T_\lambda \text{ML}$.

5.3 The symplectomorphism theorem

Fix a strict valuation $v \in \text{ML}$, and recall it identifies the set of multicurves with Λ_v^+. Let T be a real tree with a free and minimal action of $\pi_1(S)$ representing v, so that $l(\alpha) = 2v(t_\alpha)$ for all $\alpha \in \pi_1(S)$.

Lemma 32 *The distance between two branch points in T belongs to Λ_v.*

Proof This lemma can be deduced from repeated applications of Proposition 25. For instance, the distance between two disjoint axes A_γ and A_δ can be written $D = \frac{1}{2}(l(\gamma\delta) - l(\gamma) - l(\delta))$ and hence belongs to Λ_v. Instead, we may prove it as a direct consequence of a more conceptual construction for T_v using Bass–Serre theory: we refer to formula (3) in [16, Section 4.1]. $\qquad\square$

Given $\phi \in T_v\text{ML} = \text{Hom}(\Lambda_v, \mathbb{R})$, we define a corresponding $c_\phi \in \text{Hom}_\pi(\mathcal{B}(T), \mathbb{R})$ by setting $c_\phi(x, y) = \frac{1}{2}\phi(d(x, y))$, where d is the distance in T. As d is $\pi_1(S)$–invariant, c is also, and the identity $c(x, z) = c(x, y) + c(y, z)$ for y between x and z follows from the triangular equality satisfied by d. In other words, there is a well-defined map

$$(7) \qquad\qquad \Psi: T_v\text{ML} \to \text{Hom}_\pi(\mathcal{B}(T), \mathbb{R}), \quad \phi \mapsto c_\phi.$$

Proposition 33 *The map Ψ induces an isomorphism $T_v\text{ML} \simeq \text{Hom}_\pi(\mathcal{B}(T), \mathbb{R})$.*

Proof The linearity of Ψ is obvious. We first prove injectivity: suppose $c_\phi = 0$. For any nontrivial $\alpha \in \pi_1(S)$, choose a branch point x on its axis A_α so that the translation length satisfies $l(\alpha) = 2v(t_\alpha) = d(x, \alpha x)$. As $c_\phi(x, \alpha x) = \frac{1}{2}\phi(d(x, \alpha x))$ we get $\phi(v(t_\alpha)) = 0$, but Λ_v is generated by the $v(t_\alpha)$ for $\alpha \in \pi_1(S)$, so $\phi = 0$.

This suggests the construction of the inverse, but this time we think of ϕ as a map $\phi\colon \mathbb{C}(X)^*/\mathcal{O}_v^\times \to \mathbb{R}$. Given $c \in \mathrm{Hom}_\pi(\mathcal{B}(T), \mathbb{R})$, we define $\phi(t_\alpha) = c(x, \alpha x)$ for any simple curve α, where x is any branch point in A_α (by additivity of c, this does not depend on the branch point). We extend ϕ to any multicurve by linearity. Finally for any $f \in \mathbb{C}[X]^*$ we set $\phi(f) = \phi(t_\mu)$, where μ is the v–extremal multicurve in f. The point is to show that ϕ indeed belongs to $T_v\mathrm{ML}$: as it satisfies equation (6) by construction, it remains to prove that it is multiplicative.

We first show that the defining property $\phi(t_\gamma) = c(x, \gamma x)$ extends to all loops $\gamma \in \pi_1(S)$ by induction on the number of self-intersections. Suppose γ has $n > 0$ intersections. Let p be one of them and denote by α and β the two elements of $\pi_1(S, p)$ such that $\gamma = \alpha\beta$. Since v is acute, we have either $v(t_{\alpha\beta^{-1}}) < v(t_\alpha) + v(t_\beta) = v(t_{\alpha\beta})$ or $v(t_\alpha) + v(t_\beta) < v(t_{\alpha\beta^{-1}}) = v(t_{\alpha\beta})$, and we apply either case (2)(i) or case (1) of [18, Proposition 1.6] (which are unmodified in [6]).

In the first case, the axes A_α and A_β intersect along a segment xy such that both isometries push x in the direction of y, and we have $x \in A_{\alpha\beta}$. If $l(\beta) \geq d(x, y)$ then $l(\alpha\beta) = d(x, \alpha\beta x) = d(x, y) + d(y, \alpha y) + d(\alpha y, \alpha\beta x)$, whence $c(x, \alpha\beta x) = c(x, y) + c(y, \alpha y) + c(y, \beta x) = c(y, \alpha y) + c(x, \beta x)$, with $x \in A_\beta$ and $y \in A_\alpha$. If $l(\beta) \leq d(x, y)$ then $d(x, \alpha\beta x) = d(x, \beta x) + d(\beta x, \alpha\beta x)$ whence $c(x, \alpha\beta x) = c(x, \beta x) + c(z, \alpha z)$ with $z = \beta x \in A_\alpha$. Each time, the induction hypothesis applies, showing that both definitions of $\phi(t_\gamma)$ coincide.

In the second case, the axes A_α and A_β are disjoint: let xy be the geodesic joining them, and note that x also belongs to the axes of $\alpha\beta$ and $\alpha\beta^{-1}$. By the induction hypothesis, $\phi(t_{\alpha\beta^{-1}})$ is equal to $c(x, \alpha\beta^{-1}x)$. Then $d(x, \alpha\beta^{-1}x) = d(x, \alpha\beta x)$, whence $c(x, \alpha\beta^{-1}x) = c(x, \alpha\beta x)$ and $2\phi(t_{\alpha\beta}) = c(x, \alpha\beta x)$ as claimed.

To finish the proof, we must consider $f, g \in \mathbb{C}[X]$ and show that $v(fg) = v(f) + v(g)$. If μ and ν are the v–extremal multicurves of f and g, then the v–extremal multicurve of fg is that of $t_\mu t_\nu$, denoted by ξ. We must show that $\phi(t_\mu t_\nu) = \phi(t_\xi) = \phi(t_\mu) + \phi(t_\nu)$. Let us prove more generally that if $\alpha = \alpha_1 \cup \cdots \cup \alpha_n$ is a multiloop then $\phi(t_\alpha) = \phi(t_{\alpha_1}) + \cdots + \phi(t_{\alpha_n})$, reasoning by induction on the self-intersection number of α.

If the components α_j are disjoint, we may replace each one of them by its v–extremal smoothing, which remain disjoint, and the result follows from the definition of ϕ. Hence suppose that α_1 and α_2 intersect at p. Up to changing the orientation of α_2, we can suppose that $v(t_{\alpha_1\alpha_2}) = v(t_{\alpha_1}) + v(t_{\alpha_2})$. The computation in the first case at the beginning of the proof shows that $\phi(t_{\alpha_1\alpha_2}) = \phi(t_{\alpha_1}) + \phi(t_{\alpha_2})$. We also have $\phi(t_{\alpha_1\alpha_2}t_{\alpha_3}\cdots t_{\alpha_n}) = \phi(t_{\alpha_1\alpha_2}) + \phi(t_{\alpha_3}) + \cdots + \phi(t_{\alpha_n})$ by the induction hypothesis. $\qquad\square$

Theorem 34 *The isomorphism $\Psi^*\colon \mathcal{B}(T)_\pi \to T_v^*\mathrm{ML}$ preserves the symplectic form.*

Explicitly, $\Psi^*(x, \alpha x) = d_v \log t_\alpha$ for all $\alpha \in \pi_1(S)$ and any branch point $x \in A_\alpha$. Indeed for all $\phi \in T_v\mathrm{ML}$, equation (7) and Definition 19 yield

$$\Psi(\phi)(x, \alpha x) = \tfrac{1}{2}\phi(d(x, \alpha x)) = \phi(t_\alpha) = d_v \log(t_\alpha)(\phi).$$
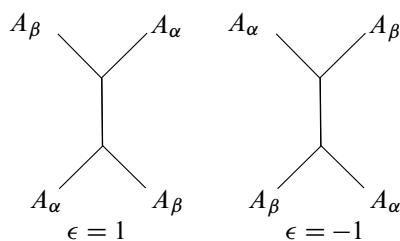
Figure 5: Sign rule for the axes.

Proof Let $\alpha, \beta \in \pi_1(S)$ represent two simple curves in S. We must prove that for $x \in A_\alpha$ and $y \in A_\beta$, $\{t_\alpha, t_\beta\}_v = \langle \pi_v, d_v \log t_\alpha \wedge d_v \log t_\beta \rangle$ equals $(x, \alpha x) \cdot_\pi (y, \beta y)$. If $i(\alpha, \beta) = 0$ then both quantities are null. Otherwise, put $\alpha \cup \beta$ in taut position.

We first compute the sum defining $\{t_\alpha, t_\beta\}_v$, in which every intersection $p \in \alpha \cap \beta$ contributes to a term $\epsilon_p(t_{\alpha_p \beta_p} - t_{\alpha_p \beta_p^{-1}})t_\alpha^{-1}t_\beta^{-1} \mod \mathcal{M}_v$. The set $\alpha \cap \beta$ is in bijection with pairs of intersecting lifts $(\widetilde{\alpha}, \widetilde{\beta}) \subset \widetilde{S} \times \widetilde{S}$ modulo the diagonal action of $\pi_1(S)$. These lifts correspond bijectively to axes of the form $(A_{\widetilde{\alpha}}, A_{\widetilde{\beta}})$ in T through the equivariant map $f: \widetilde{S} \to T$ which preserves the cyclic orientations on the boundaries. Fixing representatives $\alpha, \beta \in \pi_1(S)$, every such pair is represented by some (A_α, gA_β) for a unique $g \in \langle \alpha \rangle \backslash \pi / \langle \beta \rangle$. Using again [Proposition 25](), we can rewrite

$$(8) \qquad \{t_\alpha, t_\beta\}_v = \sum_{g \in \langle \alpha \rangle \backslash \pi / \langle \beta \rangle} \epsilon(A_\alpha, gA_\beta) = \sum_{g \in \langle \alpha \rangle \backslash \pi / \langle \beta \rangle} \epsilon(A_\alpha, A_{g\beta g^{-1}}),$$

where $\epsilon(A_\alpha, A_\beta) = \pm 1$ if A_α and A_β are as in [Figure 5](), and $\epsilon(A_\alpha, A_\beta) = 0$ in any other configuration. Notice that this formula does not depend on the orientations of the axes, but on their cyclic orders at the branch points of the tree.

To end the proof, we fix $x \in A_\alpha$ and $y \in A_\beta$ to compare formula [(8)]() with $\sum_{g \in \pi}(x, \alpha x) \cdot (gy, g\beta y)$. Grouping them depending on the class of g in $\langle \alpha \rangle \backslash \pi / \langle \beta \rangle$, we are reduced to the following equality, which is easily checked:

$$\epsilon(A_\alpha, A_\beta) = \sum_{m,n \in \mathbb{Z}} (\alpha^n x, \alpha^{n+1} x) \cdot (\beta^m y, \beta^{m+1} y). \qquad \square$$

6 Identifying the symplectic tangent models

Following [19, Section 3.2], we recall Thurston's description for the tangent space to ML at a maximal measured lamination λ. We start with an orientation covering $p: S' \to S$, which is a ramified covering of degree 2 with one ramification point in each triangle of the complement $S \setminus \lambda$, and such that the preimage $p^{-1}(\lambda)$ is naturally cooriented (meaning that its normal bundle is oriented). By the Gauss–Bonnet theorem, the set R of ramification points has $4g - 4$ elements and the monodromy of the covering is a homomorphism $\rho: \pi_1(S \setminus R) \to \{\pm 1\}$, which is nontrivial around each ramification point. For later

purposes, it will be useful to consider the orbifold S^o where ramification points are thought as conical singularities of order 2.

Let $H_1(S', \mathbb{R})^{\pm}$ be the symmetric and antisymmetric part of $H_1(S', \mathbb{R})$ with respect to the involution of the covering: they are orthogonal for the intersection form. Hence (half) the intersection form restricted to $H_1(S', \mathbb{R})^-$ is nondegenerate. We shall refer to this symplectic space as Thurston's model for $T_\lambda^* \mathrm{ML}$. We can avoid introducing the covering by considering instead the homology group $H_1(S^o, \mathbb{R}^-)$ with coefficients in the $\pi_1(S^o)$–module $\mathbb{R}$ together with the action given by $\gamma \cdot x = \rho(\gamma)x$. The twisted intersection product $H_1(S^o, \mathbb{R}^-) \times H_1(S^o, \mathbb{R}^-) \to H_0(S^o, \mathbb{R}) = \mathbb{R}$ coincides with the previous definition for Thurston's model. We will stick to this point of view in the sequel.

Let T be a trivalent real tree endowed with a free minimal action of $\pi_1(S)$, which is dual to a measured geodesic lamination λ. In the next section we first recover a model for S^o which depends only on T: our space will be an infinite dimensional CW–complex homotopic to S^o. As a consequence, its fundamental group is canonically attached to T and its homology will be easy to compute from T. We will use it extensively to prove that the Bonahon model $\mathcal{B}(T)_\pi$ and Thurston model $H_1(S^o, \mathbb{R}^-)$ are naturally isomorphic symplectic vector spaces.

6.1 A homotopical construction of the orbifold tree

6.1.1 Idea of the construction We first construct a space corresponding to the tree T with an orbifold singularity of order 2 at every branch point. As the topology of T induced by the metric is not given by a cell structure, our first task is to build a cellular model of T.

To motivate our construction, let us begin with the following analogy: suppose we wish to replace the real line $\mathbb{R}$, with its usual topology, by a CW–complex whose 0–cells consist of the set $\mathbb{Q}$ of rationals with the discrete topology. We may first add a 1–cell between every pair of distinct 0–cells to make the space connected. This creates a 1–cycle for every triple of distinct rational points, so we attach a 2–cell to each of those in order to make the space simply connected. Now every 4–tuple of rationals form the vertices of a 2–cycle, to which we attach a 3–cell, and so on. In the limit, we obtain Milnor's join construction $E\mathbb{Q}$, which is a space homotopic to $\mathbb{R}$ endowed with a free and proper action of $\mathbb{Q}$.

We shall play a similar game, replacing $\mathbb{R}$ by the real tree T, and $\mathbb{Q}$ by its set of branch points $V(T)$. We first attach a 1–cell to every pair of distinct branch points. However, we close the triangle (x, y, z) only if $x, y, z \in V(T)$ belong to a same geodesic in T. Then we go on similarly in higher dimensions, so that our space will resemble $E\mathbb{Q}$ in restriction to any geodesic of T. At this stage, we have a space on which $\pi_1(S)$ acts freely and properly. As it is contractible, its quotient by $\pi_1(S)$ is homotopic to S. Next comes the orbifold singularities: in homotopy theory, this is represented by a $K(\mathbb{Z}/2, 1)$–space, that is, $\mathbb{RP}^\infty$. It remains to blow up the preceding construction at every branch point and insert an infinite-dimensional projective space. This construction may look complicated but we shall do it in one shot and few lines below.
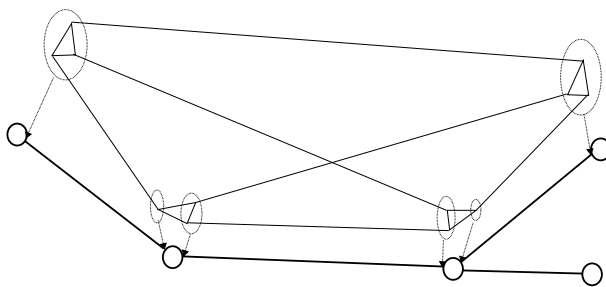
Figure 6: Attaching a 3–cell in T^o.

6.1.2 Formal construction A *half-edge* of T is a pair (x, h) consisting of a branch point x of T and a connected component h of $T \setminus \{x\}$; we sometimes just write h, as it determines x. Let us construct a CW–complex T^o whose 0–skeleton is the set of half-edges of T. First, we attach a 1–cell denoted by (h, k) between every pair of half-edges incident to the same branch point $x \in V(T)$. Now at every branch point x, the incident half-edges h_1, h_2, h_3 form a triangle homeomorphic to $\mathbb{RP}^1$, along which we attach a copy of $\mathbb{RP}^\infty$. For the moment, T^o is a disjoint union of infinite projective spaces indexed by the set of branched points $V(T)$; we call it the *orbifold part*.

Now, we add a *connecting part*, as suggested in Figure 6. Fix $\epsilon > 0$ small enough, say $\frac{1}{3}$. Consider a finite set W of branch points $\{x_0, \ldots, x_n\}$ aligned on a geodesic of T, and denote by h_i and k_i the half-edges incident to x_i containing (a nonempty) part of that geodesic. The n–cell

$$\Delta_W = \left\{ (r_x)_{x \in W} \in [0, 1-\epsilon]^W \; \middle| \; \sum_{x \in W} r_x = 1 \right\}$$

is a truncated simplex, and there is an obvious inclusion $\Delta_{W'} \subset \Delta_W$ when $W' \subset W$. The face of Δ_W truncated at x_i corresponds to the set $\Delta_W^{x_i}$ of families (r_x) satisfying $r_{x_i} = 1 - \epsilon$. We attach $\Delta_W^{x_i}$ to the orbifold part of T^o through the map $W \setminus \{x_i\} \to \{h_i, k_i\}$ sending the branch point x_j to the half-edge based at x_i which contains x_j, as in Figure 6. The 1–cells $\Delta_{\{x,y\}}$ will be called edges and denoted by (x, y).

As promised, the action of $\pi_1(S)$ on T^o is now proper, so that we may form the quotient $\Sigma^o = T^o / \pi_1(S)$. The following lemma shows that Σ^o and S^o are homotopic. Interestingly, the proof consists in constructing an equivariant map $F : T^o \to \widetilde{S}^o$, which plays the role of a (nonexistent) retraction for the map $f : \widetilde{S} \to T$.

Lemma 35 *Let $\widetilde{S}^o$ be the covering of the orbifold S^o corresponding to the kernel of the natural map $\pi_1(S^o) \to \pi_1(S)$. There exists a $\pi_1(S)$–equivariant map $F : T^o \to \widetilde{S}^o$ which induces a homotopy equivalence between Σ^o and S^o.*

Proof To define F, represent T as the dual tree to a measured geodesic lamination λ, and consider the collection of circles inscribed in each triangle of the complement $S \setminus \lambda$: they lift to a collection of

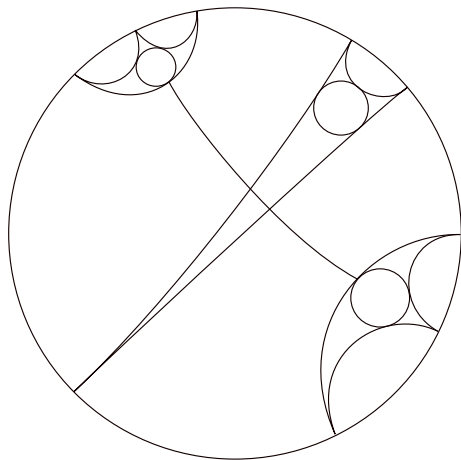Figure 7: Lifting a geodesic to $\mathbb{H}^2$ (done with Geogebra).

circles C_x in $\widetilde{S} \simeq \mathbb{H}^2$ indexed by $x \in V(T)$. Moreover, the half-edges incident to x correspond bijectively to the three intersection points of C_x with the leaves of the lamination; see Figure 7.

The covering $\widetilde{S}^o$ is obtained from $\mathbb{H}^2$ by drilling out the interior of C_x and gluing back a copy of $\mathbb{RP}^\infty$ along $\mathbb{RP}^1 \simeq C_x$ for every $x \in V(T)$. By construction, the orbifold S^o is homotopic to the quotient $\widetilde{S}^o/\pi_1(S)$.

We now proceed to the construction of an equivariant map $F: T^o \to \widetilde{S}^o$. There is already an identification between the orbifold parts of both spaces, so that we are left to define the map F on the connecting part.

For every pair $(x, y) \in V(T)^2$, we must define a path $F(x, y)$ in $\widetilde{S}^o$ connecting the points of C_x and C_y identified to the endpoints h_x, h_y of (x, y) in T^o. A first guess would be to consider the geodesic path γ between the points h_x and h_y. This path actually projects to the geodesic joining x to y in T. However, it may intersect a forbidden circle C_z, in which case it enters its circumscribed ideal triangle Δ_z by one side and leaves it by another. Call p_z the ideal vertex at the intersection of these two sides. We can homotope γ inside Δ_z to a path avoiding C_z which stays on the side containing p_z; see Figure 7.

Moreover, we can choose those paths in such a way that F is $\pi_1(S)$–equivariant. Let us now consider a triple of points x, z, y lying on a geodesic of T in that order. We have defined $F(x, z)$, $F(z, y)$ and $F(x, y)$: it is not hard to see that the region enclosed by the three arcs and the boundary of C_z does not contain any other circle, hence it can be filled by a triangle: this extends F to the 2–skeleton of T^o. This procedure can be continued to define an equivariant map $F: T^o \to \widetilde{S}^o$, which induces a map $\overline{F}: \Sigma^o \to S^o$.

We would like to show that $\overline{F}$ is a homotopy equivalence. The space $\widetilde{S}^o$ is Eilenberg–Mac Lane, and Lemma 36 below shows that so is T^o, hence it is sufficient to prove that $\overline{F}$ induces an isomorphism between fundamental groups. Behold the following commutative diagram, and observe that the five

lemma reduces the statement to showing that F_* is an isomorphism:

$$
\begin{array}{ccccccccc}
0 & \longrightarrow & \pi_1(T^o) & \longrightarrow & \pi_1(\Sigma^o) & \longrightarrow & \pi_1(S) & \longrightarrow & 0 \\
& & \downarrow{\scriptstyle F_*} & & \downarrow{\scriptstyle \bar{F}_*} & & \| & & \\
0 & \longrightarrow & \pi_1(\widetilde{S}^o) & \longrightarrow & \pi_1(S^o) & \longrightarrow & \pi_1(S) & \longrightarrow & 0
\end{array}
$$

This last statement is clear from the fact that $\pi_1(T^o)$ and $\pi_1(\widetilde{S}^o)$ are both isomorphic to a free product of copies of $\mathbb{Z}/2\mathbb{Z}$ indexed by $V(T)$; see again Lemma 36. $\qquad \square$

6.2 Homology of T^o

The homology of T^o can be computed from its finite subcomplexes, which are easy to understand thanks to the following lemma. For a finite set $W \subset V(T)$, let $T^o_{(W)}$ be the union of cells involving W only: a cell belongs to $T^o_{(W)}$ when all its 0–faces are of the form (x, h) for $x \in W$. We define T^o_W to be the subcomplex of $T^o_{(W)}$ whose connecting part reduces to the 1–cells (x, y) for $x, y \in W$ such that there is no other element in W on the geodesic joining them. In more intuitive terms, T^o_W is a collection of $\mathbb{RP}^\infty$ indexed by W, connected in a tree-like fashion given by the embedding of W in T.

Lemma 36 *For all finite $W \subset V(T)$, the cell complex $T^o_{(W)}$ retracts by deformation on T^o_W.*

Proof We define the retraction by induction on the maximal dimension of the truncated simplices $\Delta_U \subset T^o_{(W)}$. Let $U = \{x_0, \dots, x_n\}$ correspond to one of them, it is the intersection of W with a geodesic in T. We retract Δ_U by deformation onto the union of $\Delta_{U'}$ for $U' \subset U$ ranging over all subsets which do not contain both x_0 and x_n. This procedure stops when $U = \{x, y\}$ and x and y are closest neighbors in W. $\qquad \square$

We define a 1–cochain $\rho \in C^1(T^o, \{\pm 1\})$ sending every 1–cell of T^o to -1. It is a cocycle because the 2–cells of T^o, being either hexagons (orbifold part) or squares (contained in some Δ_W for W of cardinality 3), have an even number of 1–faces. The geometric idea underlying this definition is that any half-edge stands for a local coorientation of the lamination $\widetilde{\lambda}$, say pointing to the closest singular point. Following an edge e in T^o (transverse to $\widetilde{\lambda}$), we arrive at the other end with the opposite coorientation, giving $\rho(e) = -1$.

This cocycle defines a homomorphism $\rho \colon \pi_1(T^o) \to \mathbb{R}$ and we denote by $\mathbb{R}^-$ the vector space $\mathbb{R}$ with the action $\gamma.x = \rho(\gamma)x$. Our first task is to compute the homology of T^o with coefficients in $\mathbb{R}$ and $\mathbb{R}^-$.

Lemma 37 *We have $H_k(T^o, \mathbb{R}) = 0$ if $k \neq 0$, $H_k(T^o, \mathbb{R}^-) = 0$ if $k \neq 1$, and*

$$
H_1(T^o, \mathbb{R}^-) \simeq \mathcal{B}(T).
$$

Proof Observe that $T^o = \varinjlim T^o_{(W)}$ as W exhausts the finite subsets of the countable set of branch points $V(T)$ and by Lemma 36, $T^o_{(W)}$ retracts by deformation on T^o_W, thus $H_*(T^o, \mathbb{R}^\pm) = \varinjlim H_*(T^o_W, \mathbb{R}^\pm)$.

We may forget about the cocycle ρ while computing the untwisted real homology, and further retract the space T^o_W onto a wedge of infinite projective spaces. Thus $H_0(T^o_W, \mathbb{R}) = \mathbb{R}$ and $H_k(T^o_W, \mathbb{R}) = 0$ for $k > 0$, so the same holds for T^o.

We now return to the twisted homology of T^o_W. For this we consider the double cover $T'_W \to T^o_W$ corresponding to ρ and compute the untwisted homology of the total space: it splits into the ± 1–eigenspaces of the involution, which coincide with $H_*(T^o_W, \mathbb{R}^\pm)$, respectively. The space T'_W is homotopy equivalent to a graph with vertex set W, and two edges above each edge e of T^o_W connecting its endpoints with opposite orientations, as shown below:
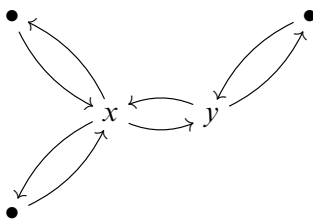
It follows that $H_0(T'_W, \mathbb{R}) = \mathbb{R}$ and $H_k(T'_W, \mathbb{R}) = 0$ if $k > 1$. Moreover $H_1(T'_W, \mathbb{R})$ has a basis formed by the cycles $c(x, y) \in H_1(T'_W, \mathbb{R})$ indexed by the edges (x, y) of T'_W, which consist in making a round trip from x to y, following the arrows. The Galois involution of T'_W exchanges the orientation of $c(x, y)$, so $H_1(T^o_W, \mathbb{R}^-)$ is freely generated by pairs (x, y) where x and y are closest neighbors in W.

Taking the limit as W converges to $V(T)$, we obtain $H_k(T^o, \mathbb{R}^-) = 0$ for $k = 0$ and $k > 1$. If an edge (x, y) gets subdivided into (x, z) and (z, y) as W increases, we have $c(x, y) = c(x, z) + c(z, y)$, which is compatible with the equality $(x, y) = (x, z) + (z, y)$, and provides the desired isomorphism for the inductive limit of $H_1(T^o_W, \mathbb{R}^-)$. $\qquad\square$

6.3 Homology of the quotient $\Sigma^o = T^o/\pi$

Let us write $\pi = \pi_1(S)$ for short. The cocycle ρ on T^o is π–invariant, so it induces a homomorphism $\pi_1(\Sigma^o) \to \{\pm 1\}$ that we also denote by ρ. The π–equivariant homotopy equivalence between Σ^o and S^o thus yields a homomorphism $\pi_1(S^o) \to \{\pm 1\}$. By the remark following Lemma 36, this homomorphism is the coorientation monodromy of λ, so its kernel corresponds to the covering $S' \to S^o$. Consequently, we may deduce the homology of S^o with coefficients in $\mathbb{R}^\pm_\rho$ from that of Σ^o with the same coefficients.

The 2–fold covering S' of S^o ramified over R satisfies $\chi(S') = 2\chi(S) - (4g - 4) = 8 - 8g$ by the Riemann–Hurwitz formula. As $H_*(S^o, \mathbb{R}^\pm) = H_*(S', \mathbb{R})^\pm$, we get that $H_*(S^o, \mathbb{R}) = H_*(S, \mathbb{R})$, whereas $H_k(S^o, \mathbb{R}^-) = 0$ if $k \neq 1$ and $\dim H_1(S^o, \mathbb{R}^-) = 6g - 6$.

On the other hand, we can compute $H_*(T^o/\pi, \mathbb{R}^{\pm})$ from $H_*(T^o, \mathbb{R}^{\pm})$ using the Cartan–Leray spectral sequence. Its second page is $E_{p,q}^2 = H_p(\pi, H_q(T^o, \mathbb{R}^{\pm}))$ and converges to $H_{p+q}(\Sigma^o, \mathbb{R}^{\pm})$. Lemma 37 implies that, with both coefficients, the second page has only one line, whence the isomorphisms

$$H_*(\Sigma^o, \mathbb{R}) = H_*(\pi, \mathbb{R}) = H_*(S, \mathbb{R}) \quad \text{and} \quad H_*(\Sigma^o, \mathbb{R}^-) = H_{*-1}(\pi, \mathcal{B}(T)).$$

This yields the proposition that we are after.

Proposition 38 *Given a maximal measured lamination λ with associated covering $S' \to S$ and corresponding tree T, there is a natural isomorphism*

$$H_1(S', \mathbb{R})^- = H_1(\Sigma^o, \mathbb{R}^-) = H_0(\pi, \mathcal{B}(T)) = \mathcal{B}(T)_\pi.$$

We also have $H_k(\pi, \mathcal{B}(T)) = 0$ for $k = 1, 2$. Observe that from Poincaré duality we get $H_2(\pi, \mathcal{B}(T)) = H^0(\pi, \mathcal{B}(T)) = \mathcal{B}(T)^\pi = 0$. It is not surprising that $\mathcal{B}(T)$ has no invariant cycles as π acts freely on $V(T)$. We do not have a similar explanation for the vanishing of $H^1(\pi, \mathcal{B}(T))$.

6.4 Intersection form

In the commutative diagram

$$\begin{array}{ccc} \widetilde{S}' & \xrightarrow{\widetilde{p}} & \widetilde{S}^o \\ \downarrow{\pi} & & \downarrow{\pi} \\ S' & \xrightarrow{p} & S^o \end{array}$$

the first column is a Galois covering of surfaces with group π. We have the identifications

$$H_1(\widetilde{S}', \mathbb{R})^- = H_1(\widetilde{S}^o, \mathbb{R}^-) = H_1(T^o, \mathbb{R}^-) = \mathcal{B}(T) \quad \text{and} \quad H_1(S', \mathbb{R})^- = H_1(S^o, \mathbb{R}^-) = \mathcal{B}(T)_\pi.$$

Proposition 39 *The isomorphisms $H_1(\widetilde{S}', \mathbb{R})^- = \mathcal{B}(T)$ and $H_1(S', \mathbb{R})^- = \mathcal{B}(T)_\pi$ preserve the symplectic forms.*

Proof Let us begin with the first isomorphism. Recall that we defined an equivariant map $F: T^o \to \widetilde{S}^o$: it sends the cell (x, y) to a path $F(x, y)$ joining the orbifold points corresponding to x and y and avoiding all other orbifold points. As the homology of the orbifold part of T^o with coefficients $\mathbb{R}^-$ vanishes identically, these paths actually define cycles in $H_1(\widetilde{S}^o, \mathbb{R}^-)$, which in $H_1(\widetilde{S}', \mathbb{R})$ are represented geometrically by $c(x, y) = \widetilde{p}^{-1}(F(x, y)) \subset \widetilde{S}'$. Notice that these cycles have a natural orientation (given by the coorientation of the lifted lamination $\widetilde{\lambda}'$).

Recalling the definition of the pairing in $\mathcal{B}(T)$ given in Proposition 30, it suffices to compute $c(x, y) \cdot c(z, t)$ in the case where (x, y) and (z, t) are disjoint or consecutive.

In the first case, the cycles $c(x, y)$ and $c(z, t)$ are also disjoint, so their intersection vanishes. In the second case, the cycles $c(x, z)$ and $c(z, y)$ only intersect in a neighborhood of z which looks like the
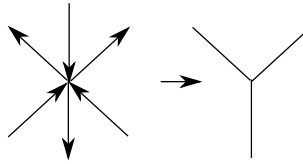
Figure 8: Double covering over a branching point.

right-hand side of Figure 8. The lifted cycles $c(x, z)$ and $c(z, y)$ go straight through the intersection point, oriented as shown. Analyzing the two possible cases, we find that the signs coincide.

Let us now consider the quotient. We showed in Section 6.3 that $H_1(S', \mathbb{R}) = H_1(\widetilde{S}', \mathbb{R})_\pi$. The result follows from the fact that the intersection form on $H_1(S', \mathbb{R})$ coincides with the averaged intersection form on $H_1(S', \mathbb{R})$.   □

References

[1]   **F Bonahon**, *Earthquakes on Riemann surfaces and on measured geodesic laminations*, Trans. Amer. Math. Soc. 330 (1992) 69–95   MR Zbl

[2]   **F Bonahon**, *Shearing hyperbolic surfaces, bending pleated surfaces and Thurston's symplectic form*, Ann. Fac. Sci. Toulouse Math. 5 (1996) 233–297   MR Zbl

[3]   **G W Brumfiel**, **H M Hilden**, SL(2) *representations of finitely presented groups*, Contemporary Mathematics 187, Amer. Math. Soc., Providence, RI (1995)   MR Zbl

[4]   **D Bullock**, *Rings of* SL$_2$(ℂ)*–characters and the Kauffman bracket skein module*, Comment. Math. Helv. 72 (1997) 521–542   MR Zbl

[5]   **D Bullock**, **C Frohman**, **J Kania-Bartoszyńska**, *Understanding the Kauffman bracket skein module*, J. Knot Theory Ramifications 8 (1999) 265–277   MR Zbl

[6]   **M J Conder**, **F Paulin**, *The translation length of the product of hyperbolic isometries of* ℝ*–trees*   MR Zbl Appendix to M J Conder, "Discrete and free two-generated subgroups of SL$_2$", J. Algebra 553 (2020) 248–267

[7]   **W M Goldman**, *The symplectic nature of fundamental groups of surfaces*, Adv. in Math. 54 (1984) 200–225   MR Zbl

[8]   **J Hass**, **P Scott**, *Shortening curves on surfaces*, Topology 33 (1994) 25–43   MR Zbl

[9]   **E Lindenstrauss**, **M Mirzakhani**, *Ergodic theory of the space of measured laminations*, Int. Math. Res. Not. 2008 (2008) art. id. rnm126   MR Zbl

[10] **F Luo**, *Simple loops on surfaces and their intersection numbers*, J. Differential Geom. 85 (2010) 73–115   MR Zbl

[11] **W Magnus**, *The uses of* 2 *by* 2 *matrices in combinatorial group theory: a survey*, Results Math. 4 (1981) 171–192   MR Zbl

[12] **J Marché**, **C-L Simon**, *Automorphisms of character varieties*, Ann. H. Lebesgue 4 (2021) 591–603   MR Zbl

[13] **M Mirzakhani**, *Counting mapping class group orbits on hyperbolic surfaces*, preprint (2016) arXiv 1601.03342

[14] **J W Morgan**, **P B Shalen**, *Valuations, trees, and degenerations of hyperbolic structures, I*, Ann. of Math. 120 (1984) 401–476 MR Zbl

[15] **J-P Otal**, *The hyperbolization theorem for fibered* 3–*manifolds*, SMF/AMS Texts and Monographs 7, Amer. Math. Soc., Providence, RI (2001) MR Zbl

[16] **J-P Otal**, *Compactification of spaces of representations after Culler, Morgan and Shalen*, from "Berkovich spaces and applications" (A Ducros, C Favre, J Nicaise, editors), Lecture Notes in Math. 2119, Springer (2015) 367–413 MR Zbl

[17] **A Papadopoulos**, **R C Penner**, *La forme symplectique de Weil–Petersson et le bord de Thurston de l'espace de Teichmüller*, C. R. Acad. Sci. Paris Sér. I Math. 312 (1991) 871–874 MR Zbl

[18] **F Paulin**, *The Gromov topology on* $\mathbb{R}$–*trees*, Topology Appl. 32 (1989) 197–221 MR Zbl

[19] **R C Penner**, **J L Harer**, *Combinatorics of train tracks*, Annals of Mathematics Studies 125, Princeton Univ. Press (1992) MR Zbl

[20] **J H Przytycki**, **A S Sikora**, *On skein algebras and* $\mathrm{Sl}_2(\mathbb{C})$–*character varieties*, Topology 39 (2000) 115–148 MR Zbl

[21] **K Rafi**, **J Souto**, *Geodesic currents and counting problems*, Geom. Funct. Anal. 29 (2019) 871–889 MR Zbl

[22] **Y Sözen**, **F Bonahon**, *The Weil–Petersson and Thurston symplectic forms*, Duke Math. J. 108 (2001) 581–597 MR Zbl

[23] **D Thurston**, *Geometric intersection of curves on surfaces*, preprint (2009) Available at https://dpthurst.pages.iu.edu/DehnCoordinates.pdf

[24] **D P Thurston**, *Positive basis for surface skein algebras*, Proc. Natl. Acad. Sci. USA 111 (2014) 9725–9732 MR Zbl

[25] **M Vaquié**, *Valuations*, from "Resolution of singularities" (H Hauser, J Lipman, F Oort, A Quirós, editors), Progr. Math. 181, Birkhäuser, Basel (2000) 539–590 MR Zbl

[26] **M Wolff**, *Connected components of the compactification of representation spaces of surface groups*, Geom. Topol. 15 (2011) 1225–1295 MR Zbl

Sorbonne Université
Paris, France

Laboratoire Paul Painlevé UMR CNRS, Université de Lille
Lille, France

julien.marche@imj-prg.fr, christopher-lloyd.simon@ens-lyon.fr

The local (co)homology theorems for equivariant bordism

Marco La Vecchia

We generalize the completion theorem for equivariant MU_G–module spectra for finite extensions of a torus to compact Lie groups using the splitting of global functors proved by Schwede. This proves a conjecture of Greenlees and May.

1 Overview

1.1 Introduction

A completion theorem establishes a close relationship between equivariant cohomology theory and its nonequivariant counterpart. It takes various forms, but in favourable cases it states that

$$(E_G^*)_{J_G}^\wedge \cong E^*(BG),$$

where E is a G–spectrum, E_G^* is the associated equivariant cohomology theory and J_G is the augmentation ideal (Definition 3.6).

The first such theorem is the Atiyah–Segal completion theorem for complex K-theory [4]. This is especially favourable because the coefficient ring $KU_G^* = R(G)[v, v^{-1}]$ is well understood, and in particular it is Noetherian, and so in this case we can view the theorem as the calculation of the cohomology of the classifying space. The good behaviour for all groups permits one to make good use of naturality in the group, and indeed [4] uses this to give a proof for all compact Lie groups G. Previous partial results were proved by Atiyah and Hirzebruch in [2; 3]. The result raised the question of what other theories enjoy a completion theorem, and the case of equivariant complex cobordism was considered soon afterwards, with Löffler giving a proof in the abelian case [16]. The fact that the coefficient ring MU_G^* is not known explicitly means that this cannot be viewed as a computation of the cohomology of the classifying space. The fact that the coefficient ring is unknown and not Noetherian was an obstacle to extensions. Despite the algebraic complexity of the coefficients, Segal made the remarkable conjecture that stable cohomotopy should satisfy the completion theorem, and this was proved for finite groups by Carlsson [6], building on important earlier work (see eg Adams, Gunawardena and Miller [1], Carlsson [5], Laitinen [13], Lin [14], Lin, Davis, Mahowald and Adams [15], Ravenel [20], Segal and Stretch [24] and Stretch [25]).

In this case, the conclusion can only be viewed as a calculation of the cohomotopy of classifying spaces in degrees 0 and below, but the structural content in positive degrees is equally striking. In the course of understanding this, there was a focus on understanding completion in various ways. From a homotopical point of view this led to the connection between completion and local cohomology and the definition of local homology (see Greenlees and May [9]), which is the derived version of completion. More precisely, for a Noetherian ring R and an ideal I, the local homology (resp. cohomology) groups $H_*^I(R; M)$ (resp. $H_I^*(R; M)$) of an R–module M calculate the left (resp. right) derived functors of completion (resp. I–power torsion); see [9] (resp. Hartshorne [12]). This derived approach led to a new proof of the Atiyah–Segal completion theorem and also its counterpart in homology; see Greenlees [8]. This in turn reopened the question of the completion theorem and local cohomology theorem for MU, but now with the challenges shifted from the formal behaviour to the algebraic behaviour: the formal structure of the proof of the local cohomology theorem for KU applies precisely for MU, but the difficulty is that, since MU_G^* is not Noetherian, it is not clear that the relevant ideals are finitely generated. Accordingly, Greenlees and May [10] isolated the formal argument and observed that if one could find a "sufficiently large" finitely generated ideal (Definition 3.8) of MU_G^*, the local cohomology and completion theorems would hold for MU. They went on to codify and use the structure of MU as a global spectrum to define and apply multiplicative norm maps, and hence construct "sufficiently large" finitely generated ideals in the case when the identity component of G is a torus. This led to the proof of local cohomology and completion theorems for MU for these groups. Much more recently, Schwede has studied global spectra more systematically [21], and in particular used the global structure of MU in a more sophisticated way to show that tautological unitary Euler classes are regular and give rise to various splittings [22].

More precisely, he proves that for every n we have a short sequence

$$0 \to MU_{U(n)}^{*-2n} \xrightarrow{e_{U(n)}(v_n)} MU_{U(n)}^* \xrightarrow{\mathrm{res}_{U(n-1)}^{U(n)}} MU_{U(n-1)}^* \to 0$$

which is split exact, and, denoting by $p_k \colon U(k) \times U(n-k) \to U(k)$ the projection to the first factor, the composite

$$MU_{U(k)}^* \xrightarrow{p_k^*} MU_{U(k) \times U(n-k)}^* \xrightarrow{\mathrm{tr}_{U(k) \times U(n-k)}^{U(n)}} MU_{U(n)}^*$$

is split injective when restricted to the kernel of the restriction map

$$MU_{U(k)}^* \to MU_{U(k-1)}^*.$$

These two facts together imply that the augmentation ideal $J_{U(n)}$ can be explicitly described as generated by the elements $s_{U(k)}^{U(n)}(e_{U(k)}(v_k))$ for $k = 1, \ldots n$ (Corollary 4.1), where $s_{U(k)}^{U(n)} \colon MU_{U(k)}^* \to MU_{U(n)}^*$ is a section of $\mathrm{res}_{U(k)}^{U(n)}$. Our contribution is to show that, for every compact Lie group G that embeds in $U(n)$, the finitely generated ideal $\mathrm{res}_G^{U(n)}(J_{U(n)}) \subset MU_G^*$ is "sufficiently large" (Corollary 5.2). This is a consequence of Schwede's results as we will see in Section 5. Working in the highly structured category of G–equivariant MU_G–modules guarantees that we can define (Definition 3.3) a homotopical version of

the stable Koszul complex for the ideal $I = \mathrm{res}_G^{U(n)}(J_{U(n)}) \subset MU_G^*$, which we denote by $K_\infty(I)$, and for an MU_G–module M we set

$$\Gamma_I(M) = K_\infty(I) \wedge_R M$$

(Definition 3.4). By a formal argument, we can then construct a morphism

$$\kappa \colon EG_+ \wedge M \to \Gamma_I(M)$$

(Construction 3.7) of G–equivariant MU_G–modules.

Theorem 1.1 *Let G be a compact Lie group with a faithful representation of dimension n, and M an MU_G–module. Then the canonical map*

$$\kappa \colon EG_+ \wedge M \to \Gamma_I(M)$$

is an equivalence of G–equivariant MU_G–module spectra.

This proves [10, Conjecture 1.4]. As a corollary, we obtain a local cohomology theorem which can be interpreted as a "derived completion theorem" for every compact Lie group.

Corollary 1.2 *Let G be a compact Lie group with a faithful representation of dimension n, M an MU_G–module, X a based G–space and $I = \mathrm{res}_G^{U(n)}(J_{U(n)}) \subset MU_G^*$. Then there are spectral sequences*

$$H_I^*(M_*^G(X)) \Rightarrow M_*^G(EG_+ \wedge X) \quad \text{and} \quad H_*^I(M_G^*(X)) \Rightarrow M_G^*(EG_+ \wedge X).$$

Since I has n generators, the local cohomology and homology are concentrated in degrees $\leq n$.

Organization

We start with a preliminary section where we introduce the notation and some basic facts of equivariant and global orthogonal spectra. In Section 3 we review the classical statement. This section is only needed to recall basic constructions and state the main theorem that we will prove in Section 5. In Section 4 we review Schwede's splitting [22, Theorem 1.4, page 5] and his corollary that ensures the regularity of certain Euler classes [22, Corollary 3.2, page 10]. Finally, in Section 5 we prove that the augmentation ideal $J_{U(n)}$ is *sufficiently large*. This will imply the completion theorem [10, Theorem 1.3, page 514] for $U(n)$ and for any compact Lie group G.

Acknowledgements

2 Notation, conventions and facts

2.1 Spaces

By a *space* we mean a *compactly generated space* as introduced in [18]. We will denote by $\mathcal{T}$ (resp. $\mathcal{T}_*$) the category of spaces (resp. pointed spaces) with continuous maps (resp. based maps). For a compact Lie group G, $\mathcal{T}_G$ denotes the category of G–spaces and G–equivariant maps. (Equivariant) mapping spaces and (based) homotopy classes of (based) maps are defined as usual, and are denoted by $\mathrm{map}(\,\cdot\,,\cdot\,)$ and $[\,\cdot\,,\cdot\,]$, respectively.

2.1.1 Universal spaces
A *family* of subgroups of a group G is a collection of subgroups closed under conjugation and taking subgroups. When G is a compact Lie group, a universal G–space for a family $\mathscr{F}$ of closed subgroups is a G–CW–complex $E\mathscr{F}$ such that:

- All isotropy groups of $E\mathscr{F}$ belong to the family $\mathscr{F}$.
- For every $H \in \mathscr{F}$ the space $E\mathscr{F}^H$ is weakly contractible.

We denote by $\widetilde{E}\mathscr{F}$ the reduced mapping cone of the collapse map $E\mathscr{F}_+ \to S^0$ which sends $E\mathscr{F}$ to the nonbasepoint of S^0. Any two such universal G–spaces are G–homotopy equivalent; hence we will refer to $E\mathscr{F}$ as the universal space for the family $\mathscr{F}$. Note that $E\{e\} = EG$.

2.2 Algebra

For a graded commutative ring A and a finitely generated ideal $I = (a_1, \ldots, a_n)$ of A, we let $K_\infty^\bullet(I)$ be the graded cochain complex

$$\bigotimes_{i=1,\ldots,n} (A \to A[1/a_i]),$$

where A and $A[1/a_i]$ sit in homological degrees 0 and 1, respectively, and the tensor product is over the ring A. If N is a graded A–module, the local cohomology groups are defined as

$$H_I^{s,t}(A; N) = H^{s,t}(K_\infty^\bullet(I) \otimes N).$$

When A is Noetherian, the functor $H_I^*(A; \cdot)$ calculates the right derived functors of the torsion functor

$$\Gamma_I(N) = \{n \in N \text{ such that } I^k n = 0 \text{ for some } k\}.$$

The main references for the theory of local cohomology are [11; 12]. Dually, we let the local homology groups be

$$H_{s,t}^I(A; N) = H_{s,t}\big(\mathrm{Hom}(\widetilde{K}_\infty^\bullet(I), N)\big),$$

where $\widetilde{K}_\infty^\bullet(I)$ is an A–free chain complex quasi-isomorphic to $K_\infty^\bullet(I)$; see [9]. When A is Noetherian and N is free or finitely generated, then the functor $H_k^I(A; \cdot)$ calculates the left derived functors of the I–adic completion functor. In particular, under these assumptions

$$H_k^I(A; N) \cong \begin{cases} N_I^\wedge & \text{if } k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

2.3 Spectra

A spectrum will be an orthogonal spectrum, and we will denote by Sp the category of orthogonal spectra as defined in [21, Definition 3.1.3, page 230]. The category Sp is closed symmetric monoidal with respect to the smash product $\cdot \wedge \cdot$. We will denote by $\mathrm{map}(\cdot, \cdot)$ the right adjoint of the smash product. Sp is also cotensored over $\mathcal{T}_*$, ie for every based space A and every spectrum X a mapping spectrum between these two is defined. We will also use the notation $\mathrm{map}(\cdot, \cdot)$ in this case. The definitions in the equivariant case are similar.

2.4 Global and equivariant stable homotopy categories

We let $\mathcal{GH}$ and $\mathcal{SH}_G$ denote respectively the *global stable homotopy category* and the G–*equivariant stable homotopy category* for a compact Lie group G. The first can be realized as a localization of the category of orthogonal spectra at the class of *global equivalences* [21, Definition 4.1.3, page 352] as constructed in [21, Theorem 4.3.18, page 400], and the second as a localization at the class of π_*–*isomorphisms* of the category of G–orthogonal spectra as constructed in [17, Theorem 4.2, page 47]. Hence,

$$\mathcal{SH}_G \cong \mathrm{Sp}_G[(\pi_*\text{–isos})^{-1}] \quad \text{and} \quad \mathcal{GH} \cong \mathrm{Sp}[(\text{global equivalences})^{-1}].$$

Both global equivalences and π_*–isomorphisms are weak equivalences of stable model structures; see [21, Theorem 4.3.17, page 398] for the global case and [17, Theorem 4.2, page 47] for the equivariant one. This implies that both categories $\mathcal{GH}$ and $\mathcal{SH}_G$ come with a preferred structure of triangulated categories, and we denote by Σ the *shift functor* in both cases. The derived smash product of Sp (resp. Sp_G) endows the category $\mathcal{GH}$ (resp. $\mathcal{SH}_G$) with a closed symmetric monoidal structure. For every compact Lie group G there is a forgetful functor $(-)_G \colon \mathcal{GH} \to \mathcal{SH}_G$ obtained from the point–set level functor of endowing a global spectrum with the trivial G–action. This functor is strong symmetric monoidal and exact; see [21, Theorem 4.5.24, page 450].

Homotopy groups are defined for equivariant spectra and for global spectra as usual [21, page 232]. In both cases, for a fixed X and k, the system of homotopy groups $\{\pi_k^H(X)\}_{H \subset G}$ has a lot of additional structure. For G–spectra, $\{\pi_k^H(X)\}_{H \subset G}$ is a *Mackey functor* [21, Definition 3.4.15, page 319]; for global spectra, $\{\pi_k^G(X)\}_{G \text{ compact Lie}}$ is a *global functor* [21, Definition 4.2.2, page 369].

If we fix a compact Lie group G and a G–spectrum X, the functor

$$\pi_k^G \colon \mathcal{SH}_G \to \mathsf{Ab}$$

is corepresented by the pair $(\Sigma^k \mathbb{S}, \mathrm{id})$, ie we have a natural isomorphism

(1) $$\mathcal{SH}_G(\Sigma^k \mathbb{S}, X) \cong \pi_k^G(X).$$

A similar statement holds in the global setting; refer to [21, Theorem 4.4.3, page 412]. Finally, for a G–spectrum X, we adopt the convention

$$X_*^G = \pi_*^G(X), \quad X_G^* = X_{-*}^G.$$

2.5 Global complex cobordism

Since our work concerns the complex bordism ring, we recollect here some facts about this theory. We will write **MU** for the global Thom ring spectrum defined in [21, Example 6.1.53] which is a model for the homotopical equivariant bordism MU_G introduced by tom Dieck in [7] for every compact Lie group G. The global theory **MU** has the structure of an *ultracommutative ring spectrum* in the sense of [21, Definition 5.1.1, page 463]; this assures that, for every compact Lie group G, the category of MU_G–modules is symmetric monoidal. For every compact Lie group G and for every unitary representation V, a *Thom class* $\sigma_G(V) \in MU_G^{2n}(S^V)$ is defined, where $n = \dim_{\mathbb{C}}(V)$. The Euler class $e_G(V) \in MU_G^{2n}$ is by definition the image of the Thom class along the fixed point inclusion $S^0 \to S^V$. If V has nontrivial G–fixed points, then the previous inclusion is G–nullhomotopic, and hence $e_G(V) = 0$ if $V^G \neq \{0\}$. On the other hand, tom Dieck showed that if $V^G = \{0\}$ then the Euler class $e_G(V)$ is a nonzero element in MU_G^{2n} [7, Corollary 3.2, page 352].

The theory **MU** has an equivariant Thom isomorphism for every unitary representation V

$$\text{(2)} \qquad\qquad MU_G^0(S^{k+V}) \cong MU_G^{-k-2n}$$

given by $\mathrm{RO}(G)$–graded multiplication with the Thom class $\sigma_G(V)$, where $n = \dim_{\mathbb{C}}(V)$. This isomorphism takes the multiplication by the class a_n defined in Section 4 to multiplication by the Euler class of the representation V.

3 Classical statement

We recall some basic constructions that can be found in [10]. To make sense of them, we need to work with highly structured equivariant ring spectra known as E_∞ ring G–spectra or commutative $\mathbb{S}_G$–algebras. In particular, all the constructions below are well defined for $R = MU_G$. We refer to [10, page 511] for a more detailed explanation and bibliography.

Construction 3.1 Let $R \in \mathcal{SH}_G$ be a G–ring spectrum as explained above. By (1), every element of R_n^G specifies by adjunction a morphism $\alpha \colon \mathbb{S} \to \Sigma^{-n} R$ in $\mathcal{SH}_G$. We let

$$\tilde{\alpha} \colon R \to \Sigma^{-n} R$$

be the composition

$$R \xrightarrow{\alpha \wedge R} \Sigma^{-n} R \wedge R \cong \Sigma^{-n}(R \wedge R) \xrightarrow{\Sigma^{-n}\mu} \Sigma^{-n} R,$$

where $\mu \colon R \wedge R \to R$ is the multiplication of R. This defines a morphism in $\mathcal{SH}_G$, and we let

$$\text{(3)} \qquad\qquad R[1/\alpha] := \text{telescope}(R \xrightarrow{\tilde{\alpha}} \Sigma^{-n} R \xrightarrow{\Sigma^{-n}\tilde{\alpha}} \Sigma^{-2n} R \to \cdots)$$

be the *mapping telescope* of the iterates of $\tilde{\alpha}$.

Remark 3.2 The *mapping telescope* in Construction 3.1 models the sequential homotopy colimit in $\mathcal{SH}_G$. For a discussion of homotopy colimits in $\mathcal{SH}_G$, refer to [19, Appendix C, page 160].

Definition 3.3 Let R and α be as above and let $I = (\alpha_1, \ldots, \alpha_n)$ be an ideal in R_*^G. We define

$$K_\infty(\alpha) := \mathrm{fib}(R \to R[1/\alpha]) \quad \text{and} \quad K_\infty(I) := K(\alpha_1) \wedge_R \cdots \wedge_R K(\alpha_n).$$

Definition 3.4 Let M be an R–module and $I \subset R_*^G$ be a finitely generated ideal. Then we define

$$\Gamma_I(M) = K_\infty(I) \wedge_R M \quad \text{and} \quad (M)_I^\wedge = \mathrm{map}_R(K_\infty(I), M).$$

Remark 3.5 There is a spectral sequence of local cohomology

$$H_I^*(R_*^G; M_*^G) \Rightarrow \Gamma_I(M)_*^G,$$

and there is a spectral sequence of local homology

$$H_*^I(R_*^G; M_*^G) \Rightarrow ((M)_I^\wedge)_G^*.$$

Note that when $M = R$ we obtain the spectral sequence that computes $K_\infty(I)_*^G$; see [10].

Definition 3.6 Let G be a compact Lie group and R be an orthogonal G–spectrum. The augmentation ideal J_G of R at G is the kernel of

$$\mathrm{res}_1^G \colon R_*^G \to R_*.$$

Construction 3.7 By construction of $R[1/\alpha]$, if $\alpha \in J_G$ then

$$\mathrm{res}_1^G R[1/\alpha] \simeq 0.$$

Hence, applying the restriction to the fibre sequence

$$\Gamma_\alpha R \to R \to R[1/\alpha],$$

we obtain a fibre sequence in which the third term is contractible. This implies, by the long exact sequence in homotopy groups induced by a fibre sequence of spectra, that the canonical map

$$\mathrm{res}_1^G(\Gamma_\alpha R) = \Gamma_{\mathrm{res}^G(\alpha)} \mathrm{res}_1^G R \xrightarrow{\simeq} \mathrm{res}_1^G R$$

is an equivalence. The same argument applies for an ideal $I \subset R_*^G$, giving an equivalence

$$\mathrm{res}_1^G(\Gamma_I R) = \Gamma_{\mathrm{res}_1^G(I)} \mathrm{res}_1^G R \xrightarrow{\simeq} \mathrm{res}_1^G R.$$

Smashing the above morphism with the universal G–space EG_+, we obtain an equivalence

$$EG_+ \wedge \Gamma_I R \to EG_+ \wedge R$$

in $\mathcal{SH}_G$. Inverting this and composing with the collapse map $EG_+ \wedge \Gamma_I R \xrightarrow{\mathrm{coll} \wedge \Gamma_I R} S^0 \wedge \Gamma_I R \cong \Gamma_I R$, we obtain a zigzag

$$EG_+ \wedge R \xleftarrow{\;\simeq\;} EG_+ \wedge \Gamma_I R \xrightarrow{\quad\overset{\kappa}{\frown}\quad} \Gamma_I R,$$

which defines a morphism of R–modules in $\mathcal{SH}_G$.

We now turn our attention to **MU** (Section 2.5) and we recall what it means for an ideal of the ring MU_G^* to be "sufficiently large". This is the key property to assure that the E^2–page of a specific spectral sequence that appears in the proof of Theorem 3.10 is zero.

Definition 3.8 [10, Definition 2.4, page 517] An ideal $I \subset MU_G^*$ is sufficiently large at H if there exists a nonzero complex representation V of H such that $V^H = \{0\}$ and the Euler class $e_H(V) \in MU_H^{2n}$ is in the radical $\sqrt{\mathrm{res}_H^G(I)}$, where $n = \dim_{\mathbb{C}}(V)$. The ideal I is sufficiently large if it is sufficiently large at all closed subgroups $H \neq 1$ of G.

Remark 3.9 Being sufficiently large is transitive with respect to subgroup inclusion, ie if $I \subset MU_*^G$ is sufficiently large then so is $\mathrm{res}_H^G(I) \subset MU_H^*$.

Theorem 3.10 [10, Theorem 2.5, page 518] *Let G be a compact Lie group. Then, for any sufficiently large finitely generated ideal $I \subset J_G$,*

$$\kappa \colon EG_+ \wedge MU_G \to \Gamma_I MU_G$$

is an equivalence in $\mathcal{SH}_G$. Therefore,

$$EG_+ \wedge M \to \Gamma_I(M) \quad \text{and} \quad (M)_I^\wedge \to \mathrm{map}(EG_+, M)$$

are equivalences for any MU_G–module M.

Proof Here, we only give a sketch of the argument, following the main reference. The point is that, if $I \subset MU_G^*$ is sufficiently large, then $\mathrm{res}_H^G I \subset MU_H^*$ is also sufficiently large. Moreover, since every descending sequence of compact Lie groups stabilizes, we can use induction and assume that the theorem holds for any proper closed subgroup of G. Passing to the cofibre of the map κ, it is enough to show that

$$\pi_*^G(\widetilde{E}G \wedge \Gamma_I MU_G) = 0.$$

We then let $\mathcal{P}$ be the family of proper subgroups of G and let $E\mathcal{P}$ be the universal space associated to $\mathcal{P}$. Since

$$\widetilde{E}\mathcal{P} \wedge S^0 \to \widetilde{E}\mathcal{P} \wedge \widetilde{E}G$$

is an equivalence, it suffices to show that $\widetilde{E}\mathcal{P} \wedge K(I)$ is contractible. Let $\mathcal{U}$ be a complete complex G–universe and define $\mathcal{U}^\perp$ to be the orthogonal complement of the G–fixed points $\mathcal{U}^G$ in $\mathcal{U}$. Then,

$$\mathrm{colim}_{V \in \mathcal{U}^\perp} S^V$$

is a model for $\widetilde{E}\mathcal{P}$. We can then compute

$$\pi_*^G(\widetilde{E}\mathcal{P} \wedge \Gamma_I MU_G) = \pi_*^G((\mathrm{colim}_{V \in \mathcal{U}^\perp} S^V) \wedge \Gamma_I MU_G) \cong \mathrm{colim}_{V \in \mathcal{U}^\perp} \pi_*^G(S^V \wedge \Gamma_I MU_G)$$

$$\cong \mathrm{colim}_{V \in \mathcal{U}^\perp} \pi_{*-|V|}^G(\Gamma_I MU_G) \cong \pi_*^G(\Gamma_I MU_G)[\{e_G(V)^{-1}\}_{V \in \mathcal{U}^\perp}].$$

Localizing the spectral sequence in Remark 3.5,

$$H_I^*(MU_*^G) \Rightarrow \pi_*^G(\Gamma_I MU_G),$$

away from the Euler classes, we obtain another spectral sequence

$$H_I^*(\boldsymbol{MU}_*^G)[\{e_G(V)^{-1}\}_{V \in \mathfrak{U}^\perp}] \Rightarrow \pi_*^G(\Gamma_I \boldsymbol{MU}_G)[\{e_G(V)^{-1}\}_{V \in \mathfrak{U}^\perp}].$$

Since local cohomology of a ring at an ideal becomes zero when localized by inverting an element in that ideal, we obtain that the E^2–term of the spectral sequence is zero for I sufficiently large. This proves the claim.   □

Remark 3.11   As stated in [10, Theorem 2.5, page 518], the previous theorem holds more generally for all commutative $\mathbb{S}_G$–algebras (or E_∞ ring G–spectra) which are equivariantly complex oriented and have natural Thom isomorphisms for unitary G–representations. For example, the theorem holds for equivariant K–theory.

The paper [10] proceeds by constructing a sufficiently large subideal of the augmentation ideal J_G whenever G is a finite group or a finite extension of a torus using "norm maps" [10, Section 3].

Here is where our approach differs from the classical one. In fact, we do not make use of norm maps, and the strategy to construct a sufficiently large subideal of J_G splits in two steps:

Step 1   We use Schwede's splitting (4) to prove that $J_{U(n)}$ is generated by "Euler classes". Thanks to this, we prove that $J_{U(n)}$ is sufficiently large (Proposition 5.1).

Step 2   Using the fact that any compact Lie group embeds into a unitary group $U(N)$ for N sufficiently large, we conclude that $\operatorname{res}_G^{U(N)} J_{U(N)}$ is a sufficiently large subideal of J_G for any compact Lie group G by Remark 3.9.

4   Schwede's splitting

We recall that a *global functor* F associates to every compact Lie group G an abelian group $F(G)$, and this association is contravariantly functorial with respect to continuous group homomorphisms. Moreover, for every closed subgroup inclusion $H < G$, a transfer map $\operatorname{tr}_H^G \colon F(H) \to F(G)$ is defined. This data needs to satisfy some relations that can be found in [21, page 373]. In [22, Theorem 1.4, page 5], Schwede proves that, for any global functor F, the restriction homomorphism

$$\operatorname{res}_{U(n-1)}^{U(n)} \colon F(U(n)) \to F(U(n-1))$$

is a split epimorphism. He then deduces a splitting of global functors when evaluated on the unitary group $U(n)$. Explicitly, the splitting takes the form

$$(4) \qquad F(U(n)) \cong F(e) \oplus \bigoplus_{k=1,\dots,n} \operatorname{Ker}\big(\operatorname{res}_{U(k-1)}^{U(k)} \colon F(U(k)) \to F(U(k-1))\big).$$

The most important application of the splitting for us is when the global functor comes from the homotopy groups of an orthogonal spectrum. In fact, for every global stable homotopy type X, that is, an object in $\mathcal{GH}$, we have a global functor

$$\underline{\pi}_*(X)(G) = \pi_*^G(X).$$

The splitting then tells us that, for every $k \leq n$, the group $\pi_*^{U(k)}(X)$ is a natural summand of $\pi_*^{U(n)}(X)$. In this case, a more explicit description of the right-hand side of the splitting is available. In fact, let ν_n be the tautological representation of $U(n)$ and let

$$a_n \in \pi_0^{U(n)}(\Sigma^\infty S^{\nu_n})$$

be the Euler class of ν_n, ie the element represented by the inclusion $S^0 \to S^{\nu_n}$. Then the short sequence

$$0 \to \pi_{*+\nu_n}^{U(n)}(X) \xrightarrow{a_n} \pi_*^{U(n)}(X) \xrightarrow{\mathrm{res}_{U(n-1)}^{U(n)}} \pi_*^{U(n-1)}(X) \to 0$$

is exact [22, Corollary 3.1, page 10].

When $X = \mathbf{MU}$, the equivariant Thom isomorphism identifies the previous short exact sequence with the following short exact sequence:

$$0 \to MU_{U(n)}^{*-2n} \xrightarrow{e_{U(n)}(\nu_n)} MU_{U(n)}^* \xrightarrow{\mathrm{res}_{U(n-1)}^{U(n)}} MU_{U(n-1)}^* \to 0.$$

Moreover, we have the following corollary:

Corollary 4.1 Let $J_{U(n)}$ be the augmentation ideal of $MU_*^{U(n)}$ (see Definition 3.6), and let $s_{U(k)}^{U(n)}$ be a section of $\mathrm{res}_{U(k)}^{U(n)}$ (see [22, Construction 1.3, page 4]). Then

$$J_{U(n)} = \big(s_{U(k)}^{U(n)}(e_{U(k)}(\nu_k)) \mid k = 1, \ldots, n\big),$$

and, in particular,

$$\mathrm{res}_{U(k)}^{U(n)} J_{U(n)} = J_{U(k)}$$

for all $k \leq n$.

Proof This is just the combination of the splitting (4) and the short exact sequence above.   □

Remark 4.2 Since the forgetful functor $(-)_G \colon \mathcal{GH} \to \mathcal{SH}_G$ is strong symmetric monoidal and exact, the global splitting (4) at the unitary group translates in a splitting in $\mathcal{SH}_G$.

Remark 4.3 Schwede [23, Definition 1.1] gives an explicit construction of the sections $s_{U(k)}^{U(n)}$, and he shows that the resulting elements

$$s_{U(1)}^{U(n)} e_{U(1)}(\nu_1), \quad s_{U(2)}^{U(n)} e_{U(2)}(\nu_2), \quad \ldots, \quad s_{U(n-1)}^{U(n)} e_{U(n-1)}(\nu_{n-1})$$

are "genuine equivariant Chern classes". In particular, they map to the classical Chern classes under the bundling map

$$MU_{U(n)}^* \to MU^*(BU(n)),$$

and they have similar naturality properties [23, Theorem 1.3].

5 The main result

Proposition 5.1 $J_{U(n)} \subseteq MU_{U(n)}$ is sufficiently large.

Proof Let H be a closed subgroup of $U(n)$. We need to show that there exists a nonzero complex H–representation with no nontrivial fixed points, the Euler class of which is in $\sqrt{\operatorname{res}_H^{U(n)}(J_{U(n)})}$. We will actually show that there is no need to take radicals in this case. Consider

$$V = \operatorname{res}_H^{U(n)} v_n - (\operatorname{res}_H^{U(n)} v_n)^H,$$

and let $k = \dim_{\mathbb{C}}(V)$. Note that $k = 0$ if and only if $H = 1$. We claim that the Euler class $e_H(V)$ is in $\operatorname{res}_H^{U(n)} J_{U(n)}$. If $k = n$, then $V = \operatorname{res}_H^{U(n)} v_n$ and

$$e_H(V) = \operatorname{res}_H^{U(n)}(e_{U(n)} v_n) \in \operatorname{res}_H^{U(n)} J_{U(n)},$$

and the claim holds.

Now let $k > 0$. We choose an orthonormal basis $(x_1, \ldots, x_{n-k})$ of $(\operatorname{res}_H^{U(n)} v_n)^H$ and a unitary matrix $g \in U(n)$ that sends the canonical basis of $\mathbb{C}^n$ to any other orthonormal basis that has as last $n-k$ vectors $(x_1, \ldots, x_{n-k})$. Then, for any $h \in H$,

$$h^g = \left(\begin{array}{c|c} \tilde{h} & 0 \\ \hline 0 & \operatorname{Id}_{n-k} \end{array} \right),$$

where $h^g = g^{-1} h g$ and $\tilde{h} \in U(k)$.

This implies that V is conjugate to the H^g–representation $\operatorname{res}_{H^g}^{U(k)}(v_k)$. Letting $g_\star \colon MU_{H^g}^* \to MU_H^*$ be the conjugation action (see the relations in [21, Definition 3.4.15, page 319]), we pass to Euler classes, obtaining the relation

$$e_H(V) = g_\star\big(e_{H^g}(\operatorname{res}_{H^g}^{U(k)}(v_k))\big).$$

We then compute the right-hand side of the last equation:

$$g_\star\big(e_{H^g}(\operatorname{res}_{H^g}^{U(k)}(v_k))\big) = g_\star\big(\operatorname{res}_{H^g}^{U(k)}(e_{U(k)}(v_k))\big) = g_\star\big(\operatorname{res}_{H^g}^{U(n)}(s_{U(k)}^{U(n)}(e_{U(k)}(v_k)))\big)$$

$$= \operatorname{res}_H^{U(n)}\big(s_{U(k)}^{U(n)}(e_{U(k)}(v_k))\big).$$

In the second equality we have used the chosen section $s_{U(k)}^{U(n)}$ (Corollary 4.1), and in the last one the formula

$$g_\star \circ \operatorname{res}_{H^g}^{U(n)} = \operatorname{res}_H^{U(n)}$$

(again, see the relations in [21, Definition 3.4.15, page 319]). By Corollary 4.1, it is clear that $\operatorname{res}_H^{U(n)}\big(s_{U(k)}^{U(n)}(e_{U(k)}(v_k))\big) \in \operatorname{res}_H^{U(n)} J_{U(n)}$, and hence we have proved the claim. Since, by construction, $e_H(V)$ is nonzero, we conclude that $J_{U(n)}$ is sufficiently large at H. $\qquad\square$

We now let G be any compact Lie group. Since every compact Lie group has a faithful representation, G is isomorphic to a closed subgroup of $U(n)$ where n is the dimension of a chosen faithful representation of G. Then we have the following corollary:

Corollary 5.2 *The ideal* $\operatorname{res}_G^{U(n)} J_{U(n)} \subset J_G$ *is a sufficiently large finitely generated ideal. Hence,* Theorem 3.10 *(the completion theorem) holds for any compact Lie group G if we choose $I = \operatorname{res}_G^{U(n)} J_{U(n)}$.*

Proof By transitivity of restrictions, the ideal $\mathrm{res}_G^{U(n)} J_{U(n)}$ is sufficiently large and is contained in J_G. The completion theorem then holds by the argument above. $\square$

Remark 5.3 Proposition 5.1 and Theorem 1.1 hold more generally for all global MU–modules.

Remark 5.4 The subideal $J = \mathrm{res}_G^{U(n)} J_{U(n)}$ of J_G is not special. Indeed, if I is any other finitely generated subideal of J_G containing J, then

$$\Gamma_I MU_G \simeq \Gamma_J MU_G \quad \text{and} \quad (MU_G)_I^\wedge \simeq (MU_G)_J^\wedge.$$

In fact, Theorem 3.10 implies that the MU_G–modules $K_\infty(I)$ and $\Gamma_I M$ are independent of the choice of I.

References

[1] **J F Adams**, **J H Gunawardena**, **H Miller**, *The Segal conjecture for elementary abelian p–groups*, Topology 24 (1985) 435–460 MR Zbl

[2] **M F Atiyah**, *Characters and cohomology of finite groups*, Inst. Hautes Études Sci. Publ. Math. 9 (1961) 23–64 MR Zbl

[3] **M F Atiyah**, **F Hirzebruch**, *Vector bundles and homogeneous spaces*, from "Differential geometry" (C B Allendoerfer, editor), Proc. Sympos. Pure Math. 3, Amer. Math. Soc., Providence, RI (1961) 7–38 MR Zbl

[4] **M F Atiyah**, **G B Segal**, *Equivariant K–theory and completion*, J. Differential Geometry 3 (1969) 1–18 MR Zbl

[5] **G Carlsson**, *G B Segal's Burnside ring conjecture for $(\mathbb{Z}/2)^k$*, Topology 22 (1983) 83–103 MR Zbl

[6] **G Carlsson**, *Equivariant stable homotopy and Segal's Burnside ring conjecture*, Ann. of Math. 120 (1984) 189–224 MR Zbl

[7] **T tom Dieck**, *Bordism of G–manifolds and integrality theorems*, Topology 9 (1970) 345–358 MR Zbl

[8] **J P C Greenlees**, *K–homology of universal spaces and local cohomology of the representation ring*, Topology 32 (1993) 295–308 MR Zbl

[9] **J P C Greenlees**, **J P May**, *Derived functors of I–adic completion and local homology*, J. Algebra 149 (1992) 438–453 MR Zbl

[10] **J P C Greenlees**, **J P May**, *Localization and completion theorems for MU–module spectra*, Ann. of Math. 146 (1997) 509–544 MR Zbl

[11] **A Grothendieck**, *Sur quelques points d'algèbre homologique*, Tohoku Math. J. 9 (1957) 119–221 MR Zbl

[12] **R Hartshorne**, *Local cohomology*, Lecture Notes in Math. 41, Springer (1967) MR Zbl

[13] **E Laitinen**, *On the Burnside ring and stable cohomotopy of a finite group*, Math. Scand. 44 (1979) 37–72 MR Zbl

[14] **W H Lin**, *On conjectures of Mahowald, Segal and Sullivan*, Math. Proc. Cambridge Philos. Soc. 87 (1980) 449–458 MR Zbl

[15] **W H Lin**, **D M Davis**, **M E Mahowald**, **J F Adams**, *Calculation of Lin's Ext groups*, Math. Proc. Cambridge Philos. Soc. 87 (1980) 459–469 MR Zbl

[16] **P Löffler**, *Bordismengruppen unitärer Torusmannigfaltigkeiten*, Manuscripta Math. 12 (1974) 307–327 MR Zbl

[17] **M A Mandell**, **J P May**, *Equivariant orthogonal spectra and S–modules*, Mem. Amer. Math. Soc. 755, Amer. Math. Soc., Providence, RI (2002) MR Zbl

[18] **M C McCord**, *Classifying spaces and infinite symmetric products*, Trans. Amer. Math. Soc. 146 (1969) 273–298 MR Zbl

[19] **T Nikolaus**, **P Scholze**, *On topological cyclic homology*, Acta Math. 221 (2018) 203–409 MR Zbl

[20] **D C Ravenel**, *The Segal conjecture for cyclic groups*, Bull. London Math. Soc. 13 (1981) 42–44 MR Zbl

[21] **S Schwede**, *Global homotopy theory*, New Math. Monogr. 34, Cambridge Univ. Press (2018) MR Zbl

[22] **S Schwede**, *Splittings of global Mackey functors and regularity of equivariant Euler classes*, Proc. Lond. Math. Soc. 125 (2022) 258–276 MR Zbl

[23] **S Schwede**, *Chern classes in equivariant bordism*, preprint (2023) arXiv 2303.12366

[24] **G B Segal**, **C T Stretch**, *Characteristic classes for permutation representations*, Math. Proc. Cambridge Philos. Soc. 90 (1981) 265–272 MR Zbl

[25] **C T Stretch**, *Stable cohomotopy and cobordism of abelian groups*, Math. Proc. Cambridge Philos. Soc. 90 (1981) 273–278 MR Zbl

Mathematics Institute, University of Warwick
Coventry, United Kingdom

marco.la-vecchia@warwick.ac.uk

Configuration spaces of disks in a strip, twisted algebras, persistence, and other stories

HANNAH ALPERT

FEDOR MANIN

We give $\mathbb{Z}$–bases for the homology and cohomology of the configuration space of n unit disks in an infinite strip of width w, first studied by Alpert, Kahle and MacPherson. We also study the way these spaces evolve both as n increases (using the framework of representation stability) and as w increases (using the framework of persistent homology). Finally, we include some results about the cup product in the cohomology and about the configuration space of unordered disks.

55R80; 16S15, 18A25, 55N31, 57Q70

1 Introduction

The configuration space of n labeled unit-diameter disks in an infinite strip of width w is denoted by config(n, w); Figure 1 depicts an example configuration. Specifically, parametrizing the configurations in terms of the centers of the disks, config(n, w) is the set of points $(x_1, y_1, \ldots, x_n, y_n) \in \mathbb{R}^{2n}$ such that $(x_i - x_j)^2 + (y_i - y_j)^2 \geq 1$ for all i and j, and such that $\frac{1}{2} \leq y_i \leq w - \frac{1}{2}$ for all i. We would like to describe the topology of config(n, w).
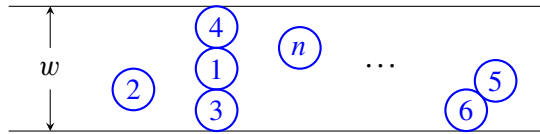
Figure 1: The configuration space config(n, w) is the set of ways to arrange n disjoint labeled disks of width 1 in $\mathbb{R} \times [0, w]$.

The topology of the configuration space of n points in the plane has been well understood since the work of Arnold [1969] and F Cohen [1976]; see [Sinha 2013] for an overview. Recently, there has been interest in more "physical" models in which the points have thickness and are constrained to lie in a bounded region, drawing inspiration from both statistical physics (as explained by Diaconis [2009]) and robotics (as explained by Farber [2008]). In the first, one imagines molecules of a substance as hard balls and extracts information about states of matter from the way they move past each other. In the latter, one can imagine a number of robots coordinating their movements so that they can travel to different points in a constrained region without bumping against each other; this amounts to a motion planning problem in a disk configuration space.

The topology of disk configuration spaces was first studied mathematically by Baryshnikov, Bubenik and Kahle [Baryshnikov et al. 2014] and experimentally by Carlsson et al. [2012]. While these papers represent real progress, trying to fully understand even the connected components of these spaces seems daunting; see [Kahle 2012]. Further work has taken two approaches to simplifying the question. One is to replace the disks by polygons, such as squares or hexagons, as in [Alpert 2020; Alpert et al. 2023]. The topology of the resulting configuration spaces is closely related to that of disk configuration spaces; in particular, it captures any of their topology that "survives for a long time" as disks grow or shrink. This observation can be formalized using persistent homology.

While it is simpler in some respects than that of disk configuration spaces, the topology of polygon configuration spaces still seems very difficult to understand. A more radical simplification, it turns out, is to remove the side walls of the rectangle, replacing it with the infinite strip, an idea introduced by Alpert, Kahle and MacPherson [Alpert et al. 2021]. That paper defines the spaces config(n, w) and computes the asymptotic growth of the rank of $H_j(\text{config}(n, w))$ as n increases, up to a constant factor depending on j and w. This turns out to be exponential unless the strip is wide compared to j.

In this paper, we present a number of results about the topology of config(n, w) and related spaces. The majority of the paper seeks to understand $H_*(\text{config}(n, w); \mathbb{Z})$, and its dependence on n and w, in greater resolution and from a more algebraic perspective. For fixed n and w, we find a geometrically motivated basis for this homology. All the classes in this basis can be assembled out of a small number of classes involving a small number of disks, depending only on w; we make this idea precise using some algebraic machinery due to Sam and Snowden [2017]. We also track the appearance and disappearance of homology classes as the width of the strip changes, using the machinery of persistent homology.

The paper includes several additional results proven using similar methods. We give a basis for the cohomology of $\text{config}(n, w)$ and make progress in understanding its cup product. We explore the possibility of FI_d–module structures on the homology of hard disk configuration spaces and no-$(k+1)$-equal spaces. Finally, we discuss the topology of the configuration space of unordered disks in a strip.

1.1 Main results

We now discuss our main results in greater detail, starting with a description of the homology of $\text{config}(n, w)$ for fixed w and all n.

Theorem A *For fixed w, $H_*(\text{config}(-, w); \mathbb{Z})$ forms a finitely generated, noncommutative twisted algebra whose generators live in $H_{\leq 3w/2-2}(\text{config}(\leq \frac{3}{2}w, w); \mathbb{Z})$.*

This contrasts with the classical family of configuration spaces of points in the plane, which forms a commutative but infinitely generated twisted algebra.

Informally, Theorem A means that $H_j(\text{config}(n, w); \mathbb{Z})$ is spanned by cycles built as follows:

(1) Separate the n disks into groups of at most $\frac{3}{2}w$.

(2) Place the groups in some order along the strip.

(3) Label the disks in some way using the numbers 1 through n.

(4) Let each group do its own thing, without interacting with the others.

The things a group can do — *elementary cycles* — come in two types: 1 to w disks can form a *wheel*, and $w + 1$ to $\frac{3}{2}w$ disks can form a *filter*. If z_1 and z_2 are elementary cycles, we refer to the act of placing them next to each other as the *concatenation product*, denoted by $z_1 \mid z_2$.

A wheel of k disks has $k - 1$ circular degrees of freedom generated by the rotation of concentric disks, making for a cycle represented by a T^{k-1}.

A filter consists of $r \geq 3$ wheels with $k > w$ disks total such that each wheel is made of at least $k - w$ disks; the wheels are ordered. The filter can move as follows. Every wheel can perform its rotations
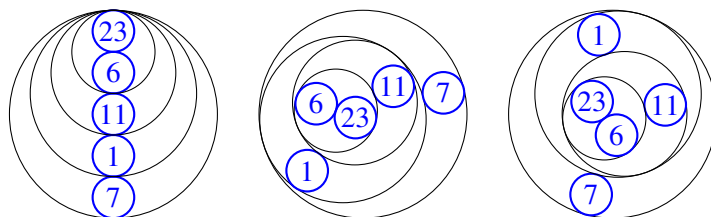

Figure 2: Some configurations of a wheel with five disks. The first configuration gives a canonical (up to switching the first two) ordering of the disks.
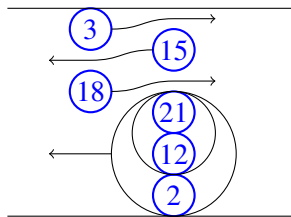
Figure 3: The wheels in a filter always cross over and under each other in the same order. This figure shows a filter with three wheels of size 1 and one wheel of size 3. The resulting cycle is an $S^2 \times T^2$.

independently, for a total of $k - r$ degrees of freedom; each wheel can also move back and forth along the strip, crossing over each other in order. Any $r - 1$ wheels can have the same x–coordinate (since they contain at most w disks total) but all r cannot. This creates an S^{r-2} inside the configuration space; thus, the whole $(k-2)$–cycle is represented by an $S^{r-2} \times T^{k-r}$.

For some purposes, it makes sense to consider filters with $r = 2$: such a filter consists of two wheels b_1 and b_2 that don't commute, and can be written as $b_2 \mid b_1 - b_1 \mid b_2$.

Our next theorem gives a basis from among the cycles generated in this way.

Theorem B $H_*(\mathrm{config}(n, w); \mathbb{Z})$ *is free abelian and has a basis consisting of concatenations of wheels and filters with* $r \geq 2$. *We say one wheel* **ranks above** *another if it has more disks, or has the same number of disks and its largest disk label is greater. A cycle is in the basis if and only if*:

   (i)   *Each wheel is ordered so that the largest label comes first.*

  (ii)   *The wheels inside each filter are in ascending order by largest label* (*regardless of the number of disks*).

 (iii)   *Adjacent wheels not inside a filter are ordered from higher to lower rank.*

 (iv)   *Every wheel immediately to the left of a filter ranks above the least wheel in the filter.*

This combinatorial structure admits a natural interpretation via homotopical algebra. Notice that $\mathrm{config}(n, w)$ naturally embeds in $\mathrm{config}(n, w + 1)$, forming a filtration. The union of this filtration is the classical configuration space $\mathrm{config}(n)$ of n points in the plane, which turns out to be homotopy equivalent to $\mathrm{config}(n, n)$. The algebraic structure of $H_*(\mathrm{config}(-))$ was considered by Arnold [1969] and Cohen [1976], who showed (in our terms) that it is a *commutative* but infinitely generated twisted algebra whose generators are wheels of all degrees.

For any two wheels in $\mathrm{config}(n)$, we have a choice of commuting them "over" or "under". However, if we order all the wheels (for example, in ascending order by largest label) and make them cross over each other in order, then they commute in a *homotopy coherent* way: informally, this means that concatenated cycles can be permuted in any sequence, or all at the same time, and that all such paths in the space of
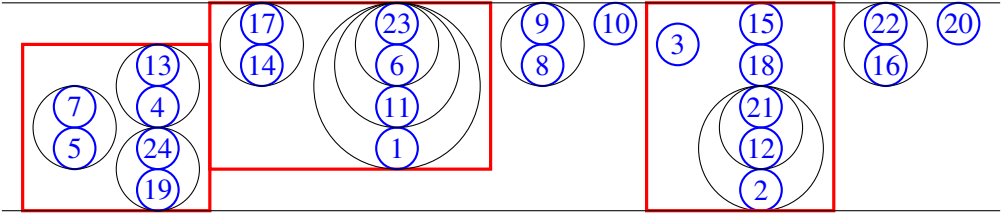
Figure 4: A basic 14–cycle in config(24, 5) represented by an $S^1 \times S^0 \times S^2 \times T^{12}$: the three red boxes are filters giving an S^1, an S^0 and an S^2 respectively, and there are 11 additional circular degrees of freedom from spinning the wheels (in black). We also remark: (a) If the disks (10) and (3) switched places, this would no longer represent a basic cycle, by property (iv) in Theorem B. (b) The single disk (10) can move freely past all the disks to its left. So, in a basic cycle, it has to appear all the way on the right.

cycles in config(n) are homotopic.[1] Just as filters with two wheels are commutators of wheels, other filters can then be thought of as nontrivial "higher commutators" that obstruct this homotopy coherence.

These higher commutators appear and disappear as we move up the filtration, increasing w, while wheels of size k are born when $w = k$ and stay forever. This observation can be made precise by considering the filtration's persistent homology.

Theorem C *The basic cycles listed in Theorem B form a $\mathbb{Z}[t]$–basis for the module $\mathrm{PH}_*(\mathrm{config}(n, *); \mathbb{Z})$. Every bar born at time w is either infinite or dies by time $2w$.*

1.2 Proof ideas

We analyze config(n, w) by relating it to a class of simpler spaces. The *no-($w{+}1$)-equal space* of n points in $\mathbb{R}$, which we denote by $\mathrm{no}_{w+1}(n, \mathbb{R})$, is the subspace of $\mathbb{R}^n$ in which at most w of the coordinates are the same. The topology of this space is fairly easy to understand, although it is related to more complicated questions about hyperplane arrangements; see [Björner and Welker 1995]. The space config(n, w) projects onto $\mathrm{no}_{w+1}(n, \mathbb{R})$ by forgetting the y–coordinates of the disks. Conversely, for every ordering on the numbered disks, there is an injective map from the no-($w{+}1$)-equal space to the configuration space of the strip in which the disks, when they meet, go around each other "in order" from top to bottom. Each subspace generated in this way is actually a retract of config(n, w).

However, in many places, two such injections coincide. For example, if we transpose two neighboring disks (a) and (b), then the two injections coincide everywhere except for an open neighborhood of the codimension-1 subspace of the no-($w{+}1$)-equal space where (a) and (b) coincide. Abstractly, we can think of this subspace as a "weighted" no-($w{+}1$)-equal space with $n - 1$ symbols, of which one has weight 2. Write $\mathrm{no}_{w+1}(n - 1, \mathcal{W})$ for this space; $\mathcal{W}$ represents the set of weights of different points.

---

[1]Unfortunately, these choices cannot be made equivariantly with respect to relabeling, which underlies the seemingly unavoidable nonequivariance of many of our results.

In the example, the subspace has codimension 1, so its neighborhood looks like $\mathrm{no}_{w+1}(n-1,\mathcal{W})\times(0,1)$; we can decompose the union of the images of the two injections into a copy of $\mathrm{no}_{w+1}(n-1,\mathcal{W})\times[0,1]$ glued onto $\mathrm{no}_{w+1}(n)$. We will see that we can model $\mathrm{config}(n,w)$ by breaking it up into layers that similarly look like thickened weighted no-$(w+1)$-equal spaces.

To compute the homology of $\mathrm{config}(n,w)$, we write it as a direct sum of homology groups of weighted no-$(w+1)$-equal spaces. We use combinatorics (specifically, discrete Morse theory) to compute the homology of these spaces.

1.3 Additional results

Besides Theorems A, B and C, the paper includes a number of other results about $\mathrm{config}(n,w)$ and related spaces.

Counting the basis elements Theorem B gives us a way to compute formulas for the Betti numbers of $\mathrm{config}(n,w)$. We describe a finite computation for each j and w that gives a formula for the rank of $H_j(\mathrm{config}(n,w);\mathbb{Z})$ as a function of n, and we show that this function is a sum of products of polynomial and exponential functions.

The cohomology of $\mathrm{config}(n,w)$ We can represent cohomology classes in $H^j(\mathrm{config}(n,w);\mathbb{Z})$ via Poincaré–Lefschetz duality as $(2n-j)$–dimensional compact submanifolds of the configuration space. We give a basis for the cohomology, showing that it is a basis by exploiting the pairing with homology. The main goal of the section is to gain some understanding of the cup product in the cohomology ring, which is fairly complicated and has many indecomposable elements — in contrast with $H^*(\mathrm{config}(n);\mathbb{Z})$, which is generated as a ring by one-dimensional classes whose pairing with homology measures the winding of two points around each other; see [Sinha 2013]. All higher-dimensional classes are linear combinations of cup products of these.

On the other hand, in $\mathrm{config}(n,w)$, pairing with cup products often cannot distinguish between $h\mid h'$ and $h'\mid h$, where h and h' are homology classes. In such cases, a cohomology class which pairs nontrivially with the commutator is perforce indecomposable. This observation allows us to prove Theorem 6.4:

Theorem *The ring $H^*(\mathrm{config}(n,w);\mathbb{Z})$ has indecomposables*

(a) *only in degree 1 when $w=2$;*

(b) *in every degree between 1 and $\lfloor\frac{1}{2}n\rfloor$ and no others when $w=3$;*

(c) *in degree 1 and in every degree between $w-1$ and $\lfloor\frac{1}{2}(n+w-3)\rfloor$ and no degree greater than $n-\lceil n/(w-1)\rceil$ when $w\geq 4$.*

Other algebraic structures One possible operation which takes j–cycles in $\mathrm{config}(n,w)$ to j–cycles in $\mathrm{config}(n+1,w)$ is inserting a singleton disk. If this operation were well defined, then $\mathrm{config}(-,w)$ would be an FI–module, one of the central objects of study in representation stability. However, since some cycles do not commute with singletons, there may be several nonequivalent ways of inserting a singleton,

separated by "barriers". In some cases, if these several ways are in turn well defined, this can be formalized via an FI_d–module structure. We show that $H_j(no_{k+1}(-, \mathbb{R}); \mathbb{Z})$ and $H_j(\mathrm{config}(-, 2); \mathbb{Z})$ have natural FI_d–module structures for appropriate d. On the other hand, for $w = 3$ we give an example which suggests that the notion of a "barrier" is not well defined and therefore there is no natural FI_d–module structure on $H_j(\mathrm{config}(-, w); \mathbb{Z})$ for $w \geq 3$.

Another strategy to pin down explicitly the algebraic structure of $H_*(\mathrm{config}(-, w); \mathbb{Z})$ is to write down a presentation for the twisted algebra using generators and relations. Category theory dictates that generators and relators in this situation are not just elements but S_n–representations for various n. We write down the generators and relations for $H_*(\mathrm{config}(-, 2); \mathbb{Z})$, but already there is some extra difficulty because of the failure of Maschke's theorem integrally. A potential avenue for future research is to write down relations for $H_*(\mathrm{config}(-, w); \mathbb{Q})$ and explore the implications for the multiplicity of various S_n–representations in $H_j(\mathrm{config}(n, w); \mathbb{Q})$.

Unordered disks   We also discuss the homology of the configuration space of unordered disks, that is, the quotient space $\mathrm{config}(n, w)/S_n$. Taking this quotient produces a large amount of torsion in the homology, so, instead of trying to write down all the torsion, we compute homology with coefficients in $\mathbb{F}_p$ and $\mathbb{Q}$. (In the bulk of the paper, we use coefficients in $\mathbb{Z}$.) The concatenation product can still be defined in this case and gives $H_*(\mathrm{config}(-, w)/S_-)$ the structure of a bigraded algebra over the base field. We compute a basis for the homology and give generators and relations for this algebra.

Structure of the paper

Section 2 contains preliminaries, including descriptions of the cell complexes we study and the algebraic framework we use in Theorem A. Sections 3 and 4 prove Theorem B, with Theorem A as a consequence; Section 3 addresses the homology of weighted no-$(k+1)$-equal spaces, and Section 4 proves its relationship to homology of configuration spaces of disks in a strip. This is the core technical content of the paper.

The rest of the sections are largely independent of each other and may appeal to different audiences (eg Section 6 to those interested in motion planning, and Section 8 to experts in representation stability). Section 5 concerns finding a formula for the Betti numbers, by counting the basis elements from Theorem B. Section 6 describes aspects of the cup product structure of the cohomology of our configuration spaces. Section 7 proves Theorem C about how the homology changes with w, the width of the strip. Section 8 concerns additional algebraic properties, in particular the question of whether our homology groups form FI_d–modules. Section 9 characterizes the homology with field coefficients in the case of unordered disks. Finally, Section 10 lists some questions for further study.

Acknowledgements

helpful conversations about this material, as well as the referee for comments that pushed us to make many sections easier to follow. Manin also thanks Shmuel Weinberger for the suggestion to look at cup products.

2 Combinatorial and algebraic setup

For the purpose of computation, Alpert et al. [2021] replace the configuration space $\mathrm{config}(n, w)$ by a homotopy-equivalent cell complex $\mathrm{cell}(n, w)$. We use the same complex; we also use the same method to find a cell complex $P(n, \mathcal{W}, k)$ that is homotopy equivalent to the weighted no-$(k+1)$-equal space $\mathrm{no}_{k+1}(n, \mathcal{W})$. In the case of ordinary, unweighted no-$(k+1)$-equal spaces this recovers a result of [Björner 2008].

In this section, we define the complexes $\mathrm{cell}(n, w)$ and $P(n, \mathcal{W}, k)$ and describe various algebraic structures on their cells which induce a similar structure on their homology. In particular, we introduce twisted algebras and show that, for each w, $H_*(\mathrm{config}(-, w))$ and $H_*(\mathrm{no}_{w+1}(-, \mathbb{R}))$ are examples. In the remainder of the paper, unless otherwise specified, homology and cohomology are always computed with coefficients in $\mathbb{Z}$.

2.1 The complex cell(n)

The cell complex $\mathrm{cell}(n, w)$ is defined as a subcomplex of a cell complex $\mathrm{cell}(n)$ described in [Blagojević and Ziegler 2014]. The complex $\mathrm{cell}(n)$ is defined in terms of the permutohedron $P(n)$, which is the $(n-1)$–dimensional polytope in $\mathbb{R}^n$ equal to the convex hull of the $n!$ points with coordinates $1, 2, \ldots, n$ in some order. The faces of $P(n)$ can be labeled by partitions of $\{1, 2, \ldots, n\}$, where the sets of the partition are ordered but the elements of each set are unordered. We refer to each set of the partition as a *block*. For instance, each vertex is labeled by a sequence of n singleton blocks, and the top-dimensional face is labeled by the one block $\{1, 2, \ldots, n\}$. Given any permutation $\sigma \in S_n$, thought of as an ordering on $\{1, 2, \ldots, n\}$, we can order the elements of each block according to σ. Then, to write out the label of a given face, we can write out the elements of each block in order, with vertical bars between blocks. We refer to a label of this form as a *symbol*. For example, the following is a symbol with four blocks that could come from the ordering $4 \prec 5 \prec 6 \prec 7 \prec 8 \prec 1 \prec 2 \prec 3$:

$$(7\,2\,|\,6\,|\,4\,5\,1\,|\,8\,3).$$

To define cell(n), we start with $n!$ copies of $P(n)$, one for each $\sigma \in S_n$. For the copy of $P(n)$ associated with σ, we use σ to label each face of that copy of $P(n)$ by a symbol. Then, whenever faces from multiple different copies of $P(n)$ have the same symbol, we identify those faces. For instance, all copies of $P(n)$ have the same vertices, but each of them has its own distinct top-dimensional face. In this way, cell(n) has exactly one cell for every possible symbol on $\{1, 2, \ldots, n\}$.
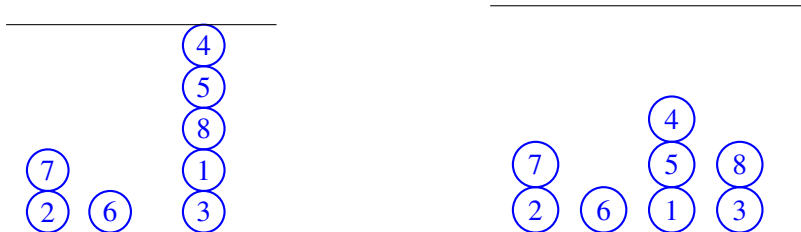
Figure 5: We can imagine each symbol of cell(n, w) as a configuration in config(n, w), where the numbers in each block are the labels in a column of disks. Pictured are configurations representing the symbol $(7\,2\,|\,6\,|\,4\,5\,8\,1\,3)$ and its face $(7\,2\,|\,6\,|\,4\,5\,1\,|\,8\,3)$.

The incidence relation on cells of cell(n) can be deduced from the geometry, or can be described explicitly on symbols as follows. A cell f in cell(n) is a face of the boundary of a cell g if g can be obtained from f by deleting a bar and *shuffling* the entries of the two neighboring blocks, preserving the ordering of the entries in each block. For example, one shuffle of $4\,6\,1\,|\,7\,3\,2$ would be $7\,4\,6\,3\,1\,2$. A d–dimensional cell consists of $n - d$ blocks.

Informally, we think of the elements of each block of a given symbol as the labels of disks in a vertical stack in config(n, w), as in Figure 5. Accordingly, as a way to specify that no more than w disks should be in each vertical stack, we define cell(n, w) to be the subcomplex of cell(n) consisting of all cells for which every block has at most w elements.

To describe the relationship between cell(n, w) and config(n, w), we start by defining config(n, w) more precisely as the set of configurations of n ordered open disks of diameter $1/w$ in the strip $\mathbb{R} \times (0, 1)$. This is a subspace of the set of configurations of n ordered distinct points in $\mathbb{R} \times (0, 1)$ such that no more than w points are on any vertical line, and [Alpert et al. 2021, Theorem 3.3] constructs a deformation retraction between the two spaces. Thus, we abuse notation and use config(n, w) to mean configurations of points with no more than w vertically aligned. We use config(n) to mean configurations of points either in $\mathbb{R} \times (0, 1)$ or in $\mathbb{R}^2$, not distinguishing between these homotopy equivalent spaces.

Theorem 2.1 *There is an affine embedding of the barycentric subdivision of* cell(n) *into* config(n) *such that, for each w, the restriction to* cell(n, w) *maps into* config(n, w) *and is a homotopy equivalence.*

Proof We use the following description of config(n) in coordinates:

$$\text{config}(n) = \{(x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{R}^n \times (0, 1)^n \mid (x_i, y_i) \neq (x_j, y_j) \text{ if } i \neq j\}.$$

To define the map on a point p in cell(n), we need to specify x–coordinates and y–coordinates, as well as check the condition $(x_i, y_i) \neq (x_j, y_j)$.

Let p be an arbitrary point in cell(n). It is in at least one of the permutohedra $P(n)$ that constitute cell(n), and $P(n)$ is embedded in $\mathbb{R}^n$, so in this sense p has coordinates in $\mathbb{R}^n$. We set the x–coordinates of the image of p to be these coordinates of p in $\mathbb{R}^n$.

For the y–coordinates, we start with the case where p is the barycenter of a cell in cell(n). To set each y_i, we find the location of the number i in the symbol of the cell of p. If i appears as the lth element of a block of size k, we set y_i to be $(k - l + 1)/(n + 1)$. In other words, for each block of size k we set the y–coordinates of the points in the block, in order, to be $k/(n + 1), (k - 1)/(n + 1), \ldots, 1/(n + 1)$. To assign y–coordinates when p is not a barycenter, we extend the map so that its restriction to each simplex of the barycentric subdivision of cell(n) is affine.

Note that, in any given cell, away from the barycenter, the y–coordinates of the points in each block remain in order. This is because, in the closure of our cell, blocks may merge but may not separate, and, when merging, the elements from each smaller block remain in order. This implies that our map is injective, because different points with the same x–coordinates are distinguished by their y–coordinates, which in each block appear in the same order as in the corresponding symbol from cell(n).

To check that the resulting map lands in config(n), suppose that we have a point p for which $x_i = x_j$. The permutohedron coordinates imply that in the symbol of the cell of p, the numbers i and j are in the same block, and so the note in the previous paragraph implies that $y_i \neq y_j$. Thus, the image of p is in config(n).

To show that the map is a homotopy equivalence, for each symbol α from cell(n), as in [Alpert et al. 2021] we define the corresponding open set $U_\alpha \subseteq$ config(n) to be the set of points $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ in $\mathbb{R}^n \times (0, 1)^n$ such that:

- Whenever i appears before j in the same block, we have $y_i > y_j$.

- Whenever i appears before j in different blocks, we have $x_i < x_j$.

- If k and l are in the same block, and k' and l' are in different blocks (with k' and l' not necessarily distinct from k and l), then we have

$$|x_k - x_l| < |x_{k'} - x_{l'}|.$$

We claim that the union of U_α where α ranges over the symbols from cell(n, w) is config(n, w). If α is a symbol of cell(n, w), then U_α is contained in config(n, w) because points from different blocks have different x–coordinates, so no more than w points can be vertically aligned. For the reverse inclusion, every element of config(n, w) is in some U_α, because we can construct α by taking the blocks to be the sets of points with the same y–coordinate, ordering the blocks from left to right, and ordering the elements of each block from top to bottom.

Each U_α is an open convex set, and [Alpert et al. 2021, Theorem 3.4] proves that the nerve of this open cover of config(n, w) is the barycentric subdivision of cell(n, w). Thus, because our map sends the barycenter of each cell α into a point of the corresponding open set U_α, our map is a homotopy equivalence for each w.   □

2.2 Signs of the boundary operator

In order to study the integral cellular chain complex of $\mathrm{cell}(n, w)$, we need to specify orientations on cells and signs for the boundary operator. To describe these, we first generalize slightly. Observe that the entries of a symbol don't have to be the numbers 1 through n, but can be any n–element set. So, for any finite set A, we have a complex $\mathrm{cell}(A)$. (Similarly, we will sometimes write $P(A)$ for the permutohedron whose coordinates are indexed by elements of A.) Moreover, there is a cellular map

$$| : \mathrm{cell}(A) \times \mathrm{cell}(B) \to \mathrm{cell}(A \sqcup B)$$

which takes a pair of symbols to the symbol obtained by putting them next to each other with a bar in between. We call this the *concatenation product*.

Any cell in $\mathrm{cell}(n)$ is either top-dimensional or a concatenation product. We define the boundary operator on a top-dimensional cell g by taking the coefficient of a cell $f = a \,|\, b$ in ∂g to be

$$(-1)^{\mathrm{length}(a)} \cdot \mathrm{sign}(\text{permutation } g \mapsto ab).$$

On a cell $g_1 \,|\, g_2$, we define ∂ via a Leibniz rule:

(2.2) $$\partial(g_1 \,|\, g_2) = \partial g_1 \,|\, g_2 + (-1)^{\dim(g_1)} g_1 \,|\, \partial g_2.$$

This defines an injective chain complex homomorphism on cellular chains,

$$| : C_*(\mathrm{cell}(A)) \otimes C_*(\mathrm{cell}(B)) \to C_*(\mathrm{cell}(A \sqcup B)),$$

using the standard tensor product on chain complexes, where the differential is defined by

$$\partial(a \otimes b) = \partial a \otimes b + (-1)^{\deg a} a \otimes \partial b.$$

Proposition 2.3   (a)   *The boundary operator defined above satisfies $\partial^2 = 0$.*

   (b)   *The S_n–action on $\mathrm{cell}(n)$ induced by permutations of $[n]$ preserves orientations of cells.*

These two features will allow us to define a twisted algebra structure on $H_*(\mathrm{cell}(-))$.

Proof   Let g be a cell in $\mathrm{cell}(n)$.

First, suppose that g is top-dimensional, and let e be a codimension-2 face of g. Then $e = e_1 \,|\, e_2 \,|\, e_3$, where e_1, e_2 and e_3 are blocks. There are two intermediate faces between e and g, which we denote by $f = b \,|\, e_3$ and $f' = e_1 \,|\, b'$. We compare

$$\mathrm{sign}(e \text{ in } \partial f) \cdot \mathrm{sign}(f \text{ in } \partial g) \quad \text{and} \quad \mathrm{sign}(e \text{ in } \partial f') \cdot \mathrm{sign}(f' \text{ in } \partial g),$$

and we show that these two products are opposite signs. This will show that $\partial^2 g = 0$.

If we consider just the contribution from the signs of the permutations, both products give the sign of the permutation relating e and g, so those contributions are equal. For the contribution from the Leibniz rule,

only the incidence between e and f' involves splitting a block that is not the first, so that incidence has a sign contribution of

$$(-1)^{\dim(e_1)} = (-1)^{\mathrm{length}(e_1)-1}$$

from the Leibniz rule, and all the other incidences have a sign contribution of 1 from the Leibniz rule. Finally, the contribution from the length of the first block gives

$$(-1)^{\mathrm{length}(b)+\mathrm{length}(e_1)} = (-1)^{\mathrm{length}(e_2)}$$

for the path through f and $(-1)^{\mathrm{length}(e_1)+\mathrm{length}(e_2)}$ for the path through f'. Taking the product of all of these, we see that the two paths give opposite signs.

If g is not top-dimensional, then g is a concatenation product $g_1 \mid g_2$, in which case we use a standard argument. The Leibniz rule gives

$$\partial^2(g_1 \mid g_2) = \partial^2 g_1 \mid g_2 + [(-1)^{\dim(g_1)} + (-1)^{\dim(\partial g_1)}] \cdot \partial g_1 \mid \partial g_2 + g_1 \mid \partial^2 g_2,$$

which is zero by induction on the number of blocks of g. This proves (a).

For (b), notice that the definitions of the signs of the boundary operator do not use any particular ordering on the numbers 1 through n. Therefore, they are invariant with respect to permutations. Since the S_n–action preserves signs of 0–cells, it preserves signs of all cells. $\qquad\square$

2.3 Twisted algebra structure

Let FB be the category of finite sets and bijective maps. Then cell$(-)$ is a functor from FB to the category of cell complexes and cellwise maps (an FB–*complex*, for short). (This is just a categorical way of saying that there is a cellular S_n–action on cell(n).) In particular, the cellular chains $C_*(\mathrm{cell}(-))$ form an FB–chain complex. Moreover, we can define a tensor product (the *Day convolution*) on FB–objects in a monoidal category by

$$(F \otimes G)(S) = \bigoplus_{A \sqcup B = S} F(A) \otimes F(B).$$

A unital monoid object with respect to this tensor product is called a *twisted algebra* (see [Sam and Snowden 2012] for a detailed discussion of twisted *commutative* algebras). In plain English, a twisted algebra is a family of objects A_n for $n = 0, 1, \ldots$ (eg abelian groups) equipped with the following structure:

- Each A_n is equipped with an S_n–action.
- For every partition of $\{1, \ldots, n\}$ into subsets of size i and j, there is a "multiplication" $A_i \otimes A_j \to A_n$. For different partitions, these multiplications commute with the S_n–action on A_n.
- There is a unit in A_0 such that multiplying by it induces the identity map on A_n.

The observations about the differential in Section 2.2 show that the concatenation product

$$|: C_i(\text{cell}(A)) \otimes C_j(\text{cell}(B)) \to C_{i+j}(\text{cell}(A \sqcup B))$$

makes $C_*(\text{cell}(-))$ into a (noncommutative) differential graded twisted algebra, or *dgta*, whose unit is the unique 0–cell ( ) in $C_*(\text{cell}(\varnothing))$. (By definition, a dgta is simply a twisted algebra in the category of chain complexes. The fact that the multiplication maps $A_i \otimes A_j \to A_n$ are chain maps forces the differential to satisfy a Leibniz rule such as (2.2).) The homology of a dgta naturally forms a graded twisted algebra. The graded twisted algebra $H_*(\text{cell}(-))$ is well understood since the work of Cohen in the 1970s and is in fact commutative; see eg [Miller and Wilson 2019], Theorem 3.4].

We are most interested in the subcomplex cell(n, w) consisting of cells whose blocks each have size at most w. Since the concatenation product of two such cells again has the same property, $C_*(\text{cell}(-, w))$ is a sub-dgta of $C_*(\text{cell}(-))$. Our goal in this paper is to understand the graded twisted algebra $H_*(\text{cell}(-, w))$, and in particular to show that it is finitely generated.

2.4 The permutohedron and no-$(k+1)$-equal spaces

The difference between cell(n) and $P(n)$ is that in cell(n) the numbers within a block are ordered and in $P(n)$ they are not. This gives a natural projection cell$(n) \to P(n)$ (forget the ordering of entries inside each block) and, for every global ordering of $1, \ldots, n$, an injective map $P(n) \to \text{cell}(n)$ (arrange the entries in each block in the given order).

$P(n)$ is a polytope, and is in particular contractible. As with cell(n), we can filter $P(n)$ by the largest size of a block, producing a sequence of complexes $P(n, k)$. These are homotopy equivalent to the no-$(k+1)$-equal space of n points in $\mathbb{R}$, as pointed out by Björner [2008, Theorem 2.4]; their homology was computed first by Björner and Welker [1995]. As with cell(n, w), $C_*(P(-, k))$ naturally has a dgta structure which induces a graded twisted algebra structure on $H_*(P(-, k))$. From the results of [Björner and Welker 1995], one sees that this is finitely generated, and in fact just has two generators: one in degree 0 (a point) and one in degree $k - 1$ (the boundary of a $P(k + 1)$). We recover this, together with a set of relations, in Section 8.

However, we are interested in a somewhat more complicated structure. Let FBW be the category of *weighted* finite sets and weight-preserving bijections. That is, every element is associated with a natural number, which is its weight. Then, given a weighted set $(A, \mathcal{W} \in \mathbb{N}^A)$, there is a complex $P(A, \mathcal{W}, k)$ which consists of all the cells for which the sum total weight of every block is at most k. The following generalization of [Björner 2008, Theorem 2.4] follows by the same argument:

Theorem 2.4 *The complex $P(A, \mathcal{W}, k)$ is homotopy equivalent to the **weighted no-$(k+1)$-equal space** $\text{no}_{k+1}(A, \mathcal{W})$ of $|A|$ points in $\mathbb{R}$ with weights $\mathcal{W}$, that is, the space of configurations of $|A|$ points in $\mathbb{R}$ such that no set of coincident points has total weight greater than k.*

Remark 2.5 The functor $C_*(P(-, k))$ is an FBW–chain complex, and in fact an FBW–dga. That is, we can define a tensor product on FBW–objects in a monoidal category by

$$(F \otimes G)(S, \mathcal{W}) = \bigoplus_{(A, \mathcal{W}_A) \sqcup (B, \mathcal{W}_B) = (S, \mathcal{W})} F(A, \mathcal{W}_A) \otimes F(B, \mathcal{W}_B),$$

and the concatenation product on $C_*(P(-, k))$ then makes it into a monoid object. This in turn makes $H_*(P(-, k))$ into a graded FBW–algebra. Once one makes all this precise, Theorem 3.4 can be interpreted as showing that this algebra is finitely generated for every k, analogously to our results about $\mathrm{cell}(-, w)$.

2.5 Generators and relations for twisted algebras

The above discussion deduces the following facts:

Theorem 2.6 *The sequences of graded abelian groups* $H_*(\mathrm{cell}(-))$, $H_*(\mathrm{cell}(-, w))$ *and* $H_*(P(-, k))$ *admit the structure of graded twisted algebras.*

To demonstrate Theorem A, we will show that wheels and filters form a finite generating set for $H_*(\mathrm{cell}(-, w))$. Later we will also give presentations of $H_*(\mathrm{cell}(-, 2))$ and $H_*(P(-, k))$ by generators and relations. To make this precise, we define a *free twisted algebra* functor F_τ from FB–modules to twisted algebras as the left adjoint to the forgetful functor U_τ from twisted algebras to FB–modules; such a functor always exists for monoids in a reasonable monoidal category [Mac Lane 1998, VII.3, Theorem 2]. Informally, a basis element for a free twisted algebra on a set $\{V_i\}$ of representations of various S_{n_i} is specified by a list of basis vectors of the various V_i, each labeled by a set of n_i labels; one takes products by concatenating these lists and retaining the labels.

If A is a twisted algebra, we say an FB–submodule $G \subset A$ *generates* A if the induced morphism $F_\tau G \to A$ is surjective; A is *finitely generated* if it has a finite-dimensional generating module G. A *presentation* of a twisted algebra A by generators and relations formally consists of a coequalizer diagram

$$F_\tau R \underset{0}{\overset{r}{\rightrightarrows}} F_\tau G \to A,$$

where G and R are FB–modules of generators and relations, respectively, and $r\colon F_\tau R \to F_\tau G$ is an FB–module homomorphism [Riehl 2017, Section 5.4]. By the adjunction, it suffices to provide a homomorphism $R \to U_\tau F_\tau G$, ie to describe the relators of A as linear combinations of words in G.

Informally, to prove finite generation, it's enough to provide a finite number of elements whose closure under multiplication and S_n–action is all of A. However, to provide a full description of a presentation, one must also understand the S_n–action on the generators. To show that A is presented by a generating set G with relations R, it is enough to show that A is generated by G and that every product of elements of G can be reduced to a basis element of A via the relators.

3 Homology of weighted no-$(k+1)$-equal spaces

In the previous section, we showed that the space $\mathrm{no}_{k+1}(n, \mathcal{W})$ retracts to a subcomplex $P(n, \mathcal{W}, k)$ of the permutohedron $P(n)$. To compute the homology of $P(n, \mathcal{W}, k)$, we use a discrete gradient vector field on $P(n)$.

3.1 Discrete Morse theory

In any cell complex, the cellular homology comes from a chain complex generated by the cells; very broadly, discrete Morse theory gives a way to decompose the chain complex as a direct sum of a chain complex that has no homology (which we discard) and a chain complex generated by a smaller subset of cells, the critical cells. To compute the homology exactly, we need to

(1) reduce to the smaller chain complex;

(2) show that the differentials in the smaller chain complex are all zero, so that the homology has a $\mathbb{Z}$–basis in bijection with the set of critical cells.

The basic definitions in discrete Morse theory are as follows. In any polyhedral cell complex, we say that cell f is a *face* of cell g if f is in the boundary of g and $\dim f = \dim g - 1$, and we say that g is a *coface* of f if f is a face of g. A *discrete vector field* on a polyhedral cell complex is a set V of pairs of cells $[f, g]$ such that f is a face of g and each cell can be in at most one pair; an example is shown in Figure 6. A discrete vector field V is *gradient* if there are no closed V–walks. A V–walk is a sequence of pairs $[f_1, g_1], \dots, [f_r, g_r]$ with $[f_i, g_i] \in V$ such that each f_{i+1} is a face of g_i other than f_i. The V–walk is closed if $f_r = f_1$.

A cell is *critical* with respect to a discrete gradient vector field V if the cell is not in any pair in V. The fundamental theorem of discrete Morse theory [Forman 2002] states that there is a cell complex that is a
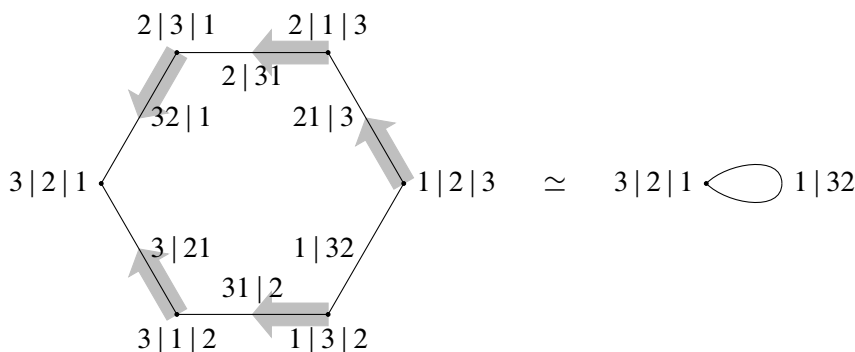

Figure 6: A *discrete gradient vector field* consists of a set of disjoint pairs of cells, each pair incident and of consecutive dimensions. The complex is homotopy equivalent to one in which the paired cells are collapsed, and only the *critical* (unpaired) cells remain.

strong deformation retraction of the original cell complex in which there is one cell per critical cell of V. Thus, we can compute the homology groups $H_j(P(n, \mathcal{W}, k))$ by defining discrete gradient vector fields and computing the homology of the collapsed chain complexes generated by the critical cells.

Our cell complexes cell(n, w) are not polyhedral cell complexes because, for instance, there are distinct cells with the same boundary. However, discrete Morse theory still gives an isomorphism on homology between the original chain complex and the collapsed chain complex as long as the following property holds: for each pair $[f, g]$ of the discrete gradient vector field, the coefficient of f in the boundary of g is a unit in our coefficient ring, in this case ± 1 because we are using coefficients in $\mathbb{Z}$. In cell(n, w), every coefficient in every boundary map is ± 1, so this property holds automatically. Alternatively, the barycentric subdivision of cell(n, w) is a polyhedral cell complex, so our arguments could be adapted to work with the barycentric subdivision instead. Thus, we proceed with the discrete Morse theory as if cell(n, w) were a polyhedral cell complex.

One way to define a discrete gradient vector field on a cell complex is by defining a total ordering on all the cells. Given a total ordering, the resulting vector field contains a pair $[f, g]$ if and only if both f is the greatest face of g and g is the least coface of f (and, for nonpolyhedral complexes, we also require the coefficient of f in the boundary of g to be a unit). One can prove that this vector field is gradient (see [Bauer 2021, Remark 3.7]).

One advantage of doing the construction in this way is that there is a simple criterion guaranteeing that the discrete gradient vector field forms a perfect Morse function. The term "perfect Morse function" refers to the case where, on the collapsed chain complex generated by the critical cells, the differential is zero, so the critical cells form a basis for the homology of the complex. The following general lemma shows that it suffices to construct, for each critical cell e, a cycle $z(e)$ such that e is its maximum cell and has coefficient equal to a unit in the coefficient ring. It turns out that such cycles automatically represent linearly independent homology classes. In our main theorems we use the lemma with $\mathbb{Z}$ coefficients, and in Section 9 we use it with field coefficients $\mathbb{Q}$ and $\mathbb{F}_p$.

Lemma 3.1 *Let X be any finite cell complex with a total ordering on the cells, giving a discrete gradient vector field. Let R be a ring of coefficients. Suppose there is a collection of cellular cycles with the following properties:*

- *The cycles in our collection are in bijection with the critical cells of the discrete gradient vector field. For each critical cell e, we denote the corresponding cycle by $z(e)$.*

- *Under the total ordering, the greatest of all the cells appearing with nonzero coefficient in the cellular chain $z(e)$ is the cell e.*

- *The coefficient of e in the chain $z(e)$ is a unit in R.*

Then the homology classes of the cycles $z(e)$ form an R–basis for $H_*(X; R)$.

Proof For any pair $[f, g]$ in the discrete vector field, we refer to f as a "match-up cell" and refer to g as a "match-down cell". We also define $z'(f)$ to be the boundary of g; we know that f is the greatest cell appearing in $z'(f)$, and that it has unit coefficient because of how we have defined the discrete vector field from the ordering.

First, we show that every j–cycle z is an R–linear combination of cycles $z(e)$ and $z'(f)$, where e ranges over the critical j–cells and f ranges over the match-up j–cells. This follows from the following observation: if a match-down cell g is the greatest cell in a j–chain, then, in the boundary of that chain, the corresponding match-up cell f appears with nonzero coefficient, because g is the least coface of f, so no other cell in the chain has f as a face. Thus, for any j–cycle z, the greatest cell of z cannot be a match-down cell. It is either a critical cell e or a match-up cell f, so we subtract the appropriate multiple of $z(e)$ or $z'(f)$ to get a new cycle with lesser maximum. Repeating this process gives us z as a linear combination of cycles $z(e)$ and $z'(f)$, so, because each $z'(f)$ is a boundary, this implies that z is homologous to a linear combination of the cycles $z(e)$ only.

To show the uniqueness, we need to show that no nontrivial linear combination of cycles $z(e)$ is null-homologous. Because the cycles $z(e)$ and $z'(f)$ have distinct maxima, they are linearly independent. Thus, it suffices to show that, if a j–cycle z is a boundary, it is a linear combination of the boundaries $z'(f)$. To see this, we look at the set of all $(j+1)$–chains. The chains $z(e)$, $z'(f)$ and g (as e ranges over all critical $(j+1)$–cells, f ranges over all match-up $(j+1)$–cells, and g ranges over all match-up $(j+1)$–cells) form an R–basis for the set of all $(j+1)$–chains, because they have distinct maxima equal to the set of all j–cells. When we apply the boundary map to this basis, the cycles $z(e)$ and $z'(f)$ map to zero, and the match-down cells g map to the j–dimensional boundaries $z'(f)$. Thus, indeed, every j–dimensional boundary is a linear combination of these boundaries $z'(f)$.

Thus, every homology class in $H_*(X)$ can be written as an R–linear combination of the homology classes of the cycles $z(e)$, and the combination is unique. $\square$

3.2 Discrete gradient vector fields on permutohedra

In what follows, we define a total ordering on all of $P(n)$, the polyhedral complex that contains $P(n, \mathcal{W}, k)$ as a subcomplex. We use the resulting discrete gradient vector fields to compute the homology, by analyzing the critical cells and constructing cycles dual to each one.

Recall that the cells of $P(n)$ are in bijection with ordered partitions of $[n]$ into blocks. We say the *weight* of a block is the sum of the weights of its elements. A block is a *singleton* if it only has one element. We assign some blocks to *leader–follower pairs* by walking left to right. A block is a *follower* if it follows a singleton (its *leader*) whose element is smaller than any of the follower's elements and which is not itself a follower. Then a total ordering $\prec$ on cells with symbols f and g is given by looking at the first block at which they differ. Let's call the two blocks f_i and g_i. Then the ordering is given as follows:

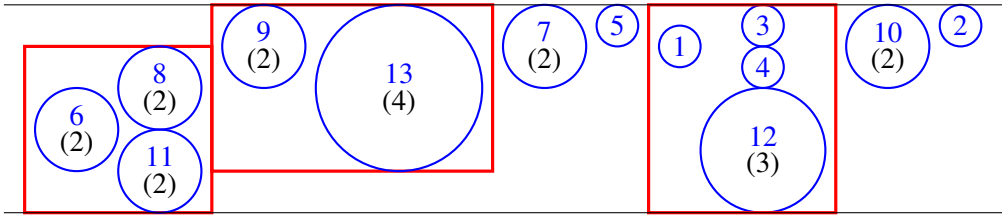(i) If f_i is a follower and g_i is not, then $f \prec g$.

Figure 7: The 3–cycle in $\mathrm{no}_6(13, \mathcal{W})$, with weights indicated by circle size and in parentheses, corresponding to the critical cell $(6 \mid 8\ 11 \mid 9 \mid 13 \mid 7 \mid 5 \mid 1 \mid 3\ 4\ 12 \mid 10 \mid 2)$. The red boxes enclose leader–follower pairs and indicate the boundaries of a $P(3)$, a $P(2)$, and a $P(4)$, respectively; thus, the cycle as a whole is represented by an $S^1 \times S^0 \times S^2$.

(ii) If f_i and g_i are both followers, then they are ordered by number of elements, and then arbitrarily (eg in lexicographic order) if they have the same number of elements.

(iii) If f_i and g_i are not followers, then we order their elements from largest to smallest; they are then ordered lexicographically, but according to a backwards order on the "alphabet". That is,

$$3 \prec 3\,2 \prec 3\,2\,1 \prec 3\,1 \prec 2 \prec 2\,1 \prec 1.$$

Lemma 3.2   *A cell in $P(n, \mathcal{W}, k)$ is critical if and only if every block is either*

(a) *a singleton which is not a follower, or*

(b) *a follower such that its weight and that of its leader add up to at least $k + 1$.*

Proof   We claim that the matching of the other cells is as follows. We look at the first block where a given cell does not match the characterization of critical given above. There are two possibilities:

- The block is a follower and, together with its leader, has total weight at most k. In that case, the cell matches *up* to the cell in which the follower and its leader are combined.

- The block is not a follower or a singleton; if preceded by a nonfollower singleton, it has at least one element smaller than that singleton. Then the cell matches *down* to the cell where the least element of the block is split off into its own block which comes before the rest. This is now a leader–follower pair, and its leader is not also a follower.

If f matches up to g, then g matches down to f, because the new combined block cannot be a follower (otherwise the leader in f would also be a follower, contradicting the definition of leader). Similarly, if g matches down to f, then f matches up to g. Thus, the match-up and match-down cells pair to form a discrete vector field.

It remains to show that this discrete vector field is induced by our ordering. Let f be a match-down cell and g be the corresponding match-up cell. We need to show f is the greatest face of g and g is the least coface of f.

To show that f is the greatest face of g, consider the result of splitting any earlier block of g. Because g looks like a critical cell at that stage, that block is a follower; after splitting, it is shorter and is still a follower, so it is smaller by property (ii) of the ordering. In contrast, among ways to split the kth block another way, or to split a later block, f is the greatest because it is the only one for which the kth block begins with the least element of the kth block of g (here we apply properties (i) and (iii)).

To show that g is the least coface of f, consider the result of combining any earlier blocks of f. Because f looks like a critical cell at that stage, the two blocks would be nonfollower singletons in decreasing order, so the combined block would be larger, not be a follower, and have the same first element as the first of the two singletons; therefore, the new block is larger by property (iii). In contrast, among ways to combine later blocks of f, g is the least because it is the only one that increases the first element of the kth block (again applying property (iii)). $\qquad\square$

3.3 Basis for homology

We now give a basis for the homology of $P(A, \mathcal{W}, k)$ for an n–element set A, thereby demonstrating that it is free abelian. By Lemma 3.1, it is sufficient to exhibit a cycle for every critical cell such that the critical cell is the largest cell in the cycle. The construction implicitly relies on an ordering on A; for it to work, we must fix an ordering so that $\mathcal{W} = (w_1, \dots, w_n)$ satisfies

(3.3) $$w_1 \leq w_2 \leq \cdots \leq w_n.$$

Every critical cell e is a concatenation product of nonfollower singletons and leader–follower pairs $(b_i \,|\, b_{i+1})$. We now build a corresponding cycle $z(e) \in P(A, \mathcal{W}, k)$. For a cell consisting of one singleton, the cycle will simply be the corresponding 0–cell. Now suppose e consists of a single leader–follower pair $(b_1 \,|\, b_2)$. Then our cycle $z(e)$ will be the boundary of the top cell $(b_1 b_2)$ of $P(A)$. Since by (3.3) each element of b_2 has greater or equal weight to the singleton b_1, every cell in the boundary is a cell of $P(A, \mathcal{W}, k)$. Moreover, $(b_1 \,|\, b_2)$ is the highest cell of $z(e)$ according to our total ordering, since b_1 is the highest-ranked block.

For a general critical cell, we define z by requiring that

$$z(e) = z(e_1 \,|\, e_2) = z(e_1) \,|\, z(e_2)$$

for any splitting $e = e_1 \,|\, e_2$ which does not split a leader–follower pair. Here $z(e_i)$ is defined based on the ordering on $A_i \subset A$ inherited from A. To see that e is the highest cell in the resulting cycle, note that the ordering on cells depends on the first block in which two cells differ. For any two cells in $z(e)$, this block will be a leader and the above argument will apply.

It is clear also that all the nonzero coefficients of $z(e)$ are ± 1. We have now proved:

Theorem 3.4 *Suppose that $\mathcal{W} = (w_1, \dots, w_n)$ is a nondecreasing sequence of weights. Then the homology $H_*(P(n, \mathcal{W}, k))$ is free abelian, with a basis given by the classes of the cycles $z(e)$, where e ranges over all cells whose blocks are of the following two types:*

(a) *a singleton which is not a follower;*

(b) *a follower such that its weight and that of its leader add up to at least $k + 1$.*

Translating these cellular cycles into cycles in $\mathrm{no}_{k+1}(n, \mathcal{W})$, we get the following picture:

- Singletons and leader–follower pairs correspond to points and groups of points arranged in order along the line.

- Every leader–follower pair corresponds to a set of r points moving back and forth of which any $r - 1$ can coincide, but all r cannot.

4 Decomposing cell(n, w) into layers

In this section we will prove Theorems A and B by expressing $H_*(\mathrm{cell}(n, w))$ in terms of the homology of weighted no-$(w{+}1)$-equal spaces. To this end, we assign an ordering to the top-dimensional cells of cell(n). For a lower-dimensional cell, its *layer* will be indexed by the first top-dimensional cell containing it; this is when that cell appears in the complex when we think of the complex as glued, in order, out of its top-dimensional cells. Recall that these top cells are identified with permutations in S_n.

The intersections of layers of cell(n) with cell(n, w) form the layers of cell(n, w). We will show that $H_*(\mathrm{cell}(n, w))$ is a direct sum of pieces which appear once each subsequent layer is glued on. Topologically, the layer associated to a permutation $\sigma \in S_n$ is a copy of $P(n - r, \mathcal{W}, w) \times [0, 1]^r$, where r and $\mathcal{W}$ depend on σ, which is glued on along $P(n - r, \mathcal{W}, w) \times \partial[0, 1]^r$; the combinatorial structure is rather more complicated. It follows (once homological triviality of the gluing is established) that the added summand of $H_j(\mathrm{cell}(n, w))$ is in bijection with elements of $H_{j-r}(P(n - r, \mathcal{W}, w))$. The main technical result of this section states:

Theorem 4.1 *The homology of* cell(n, w) *decomposes as*

$$H_*(\mathrm{cell}(n, w)) = \bigoplus_{\sigma \in S_n} H_{*-\#\sigma}\big(P(n - \#\sigma, \mathcal{W}(\sigma), w)\big).$$

We will define $\#\sigma$ and $\mathcal{W}(\sigma)$ combinatorially.

From the configuration space point of view, the new cycles in layer σ are those basic cycles from Theorem B that have a particular collection of wheels (irrespective of how those wheels are grouped into filters). The bijection above is given by replacing wheels of k disks by points of weight k. Thus, for example, this bijection in an appropriate layer takes the 14–cycle depicted in Figure 4 to the 3–cycle depicted in Figure 7.

The decomposition depends in a crucial way on the ordering of labels of disks; it is not at all equivariant with respect to the S_n–action on cell(n, w). Therefore, the methods of this section will tell us little about the S_n–module structure on the homology.

4.1 Combinatorial description of layers

Given a cell in cell(n) broken up into blocks, we further (deterministically) break up each block into *wheels*: each entry of a block is the *axle* of a wheel if it is the largest entry of the block up to that point, and the wheel consists of the axle and all the following smaller entries before the next axle.

Proposition 4.2 *Given a symbol f, the following represent the same permutation $\sigma(f)$ of $[n]$:*

(i) *The lexicographically least **shuffle** of f. A shuffle is a permutation in which the order of every block is preserved.*

(ii) *The permutation obtained by arranging all the wheels, regardless of block, in ascending order by axle.*

Proof The first number in the lexicographically least shuffle is an axle, because the first element of every block is an axle. Thus, it must be the least axle. The remaining elements of the wheel of that axle are less than that axle and thus are also less than all the other axles. Thus, the next numbers in the lexicographically least shuffle are the remaining elements of the first wheel. Repeating the same argument for each wheel, we deduce inductively from left to right that the two permutations are identical. □

For example, the symbol

$$f = (7\,2\,|\,6\,|\,4\,5\,8\,1\,3)$$

has five wheels: 7 2, 6, 4, 5 and 8 1 3; and therefore $\sigma(f) = 4\,5\,6\,7\,2\,8\,1\,3$. We say that f is in the *layer* $L(\sigma(f))$ of cell(n), or of cell(n, w). By Proposition 4.2, being in a given layer is equivalent to having a given set of wheels. We also write

$$\#\sigma = n - \text{the number of wheels of } \sigma.$$

Notice that $\#\sigma$ is the dimension of the lowest-dimensional cells of $L(\sigma)$, in which every wheel is its own block.

Let g be a boundary cell of f. Then one of the following holds:

(1) The splitting of a block to make g respects the wheels of f: each wheel goes completely into one of the new blocks. Then $g \in L(\sigma(f))$.

(2) At least one wheel of f is decomposed into two or more wheels of g. Then $\sigma(g) \prec \sigma(f)$ lexicographically.

Thus, we can build up cell(n) or cell(n, w) by gluing each subsequent layer, in lexicographical order, onto the union of the previous ones. In other words, the layers define a filtration by subcomplexes

$$L(\leq \sigma) = \bigcup_{\tau \leq \sigma} L(\tau).$$

Now, for each cell of $L(\sigma)$, we can obtain a cell of $P(\{\text{wheels of } \sigma\})$ with the same block structure, replacing each wheel by a single label.

Proposition 4.3 *The k–cells of the layer $L(\sigma)$ of cell(n, w) are in incidence-preserving bijection with the $(k-\#\sigma)$–cells of the complex $P(n - \#\sigma, \mathcal{W}(\sigma), w)$, where $\mathcal{W}(\sigma)$ consists of the cardinalities of the wheels of σ. In particular, the cells of the layer $L(\sigma)$ of cell(n) are in incidence-preserving bijection with those of $P(n - \#\sigma)$.*

Proof All cells of $L(\sigma)$ have exactly the same wheels, and the elements of each wheel always appear consecutively. Thus, given a symbol in $L(\sigma)$, we can view it as a sequence of these wheels, separated by vertical bars between blocks. The resulting new symbol is a symbol of $P(n - \#\sigma, \mathcal{W}(\sigma), w)$, because instead of n numbers in each symbol we have $n - \#\sigma$ wheels. If the cell of the original symbol has dimension k, it has $n - 1 - k$ bars, so the cell of the new symbol also has $n - 1 - k$ bars, and thus has dimension $n - \#\sigma - 1 - (n - 1 - k) = k - \#\sigma$.

As in option (1) above, cell g in $L(\sigma)$ is a boundary cell of f in $L(\sigma)$ if and only if a block in f is split to form g such that each wheel in the block is assigned entirely to the left or entirely to the right. Reinterpreting the symbols to be sequences of wheels, this is the same as the criterion for incidence in $P(n - \#\sigma, \mathcal{W}(\sigma), w)$. □

In fact, this bijection is not just combinatorial, but can be understood from several points of view: via cellular maps, cellular chains or configurations. In the next subsection, we will construct an injective cellular map

$$[0, 1]^{\#\sigma} \times P(n - \#\sigma, \mathcal{W}(\sigma), w) \to \text{cell}(n, w)$$

which matches each product of $[0, 1]^{\#\sigma}$ with an $(k-\#\sigma)$–cell to a k–cell in $L(\sigma)$. In particular, the image of $\{\frac{1}{2}\}^{\#\sigma} \times P(n - \#\sigma, \mathcal{W}(\sigma), w)$ forms a "core" or "spine" inside the layer, as illustrated in Figure 8. In

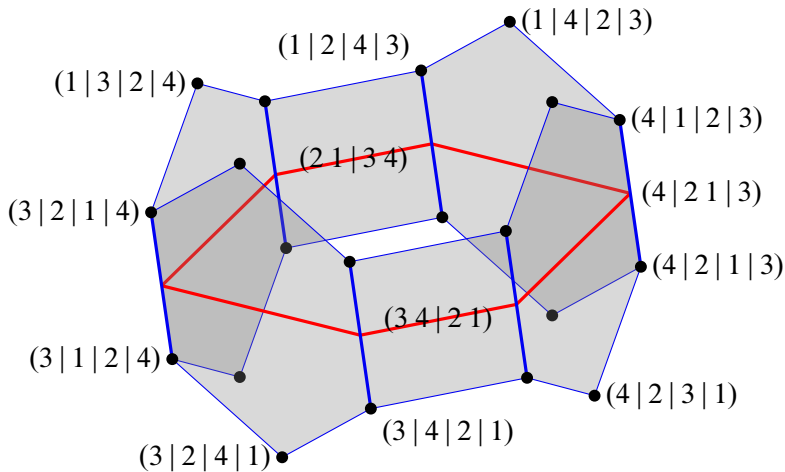

Figure 8: The layer of cell$(4, 3)$ corresponding to the permutation $\sigma = 2\ 1\ 3\ 4$. Selected cells are labeled, and the core (a copy of $P(3, \mathcal{W}(\sigma), 3)$, which is the boundary of a $P(3)$) is highlighted in red. The image of $\text{spin}_{\sigma,3}$ is combinatorially the double of the layer along the top and bottom: in the additional cells, $2\ 1$ is replaced by $1\ 2$.

fact, this will be the restriction of a (likewise injective and cellular) map

$$\mathrm{spin}_{\sigma,w} : T^{\#\sigma} \times P(n - \#\sigma, \mathcal{W}(\sigma), w) \to L(\leq \sigma)$$

to one top cell of the torus. Here $T^{\#\sigma}$ is given a product cell structure induced by a cell structure on S^1 with two vertices and two edges. In turn, $\mathrm{spin}_{\sigma,w}$ is the restriction of a map

$$\mathrm{spin}_\sigma : T^{\#\sigma} \times P(n - \#\sigma) \to \mathrm{cell}(n).$$

From the point of view of cellular chains, $\mathrm{spin}_{\sigma,w}$ induces a chain map

$$i_\sigma : C_{*-\#\sigma}\big(P(n - \#\sigma, \mathcal{W}(\sigma), w)\big) \to C_*(L(\leq \sigma))$$

via

$$i_\sigma(z) = \mathrm{spin}_{\sigma,w}([T^{\#\sigma}] \times z).$$

In the other direction, we get a surjective chain map

$$p_\sigma : C_*(L(\leq \sigma)) \to C_{*-\#\sigma}\big(P(n - \#\sigma, \mathcal{W}(\sigma), w)\big)$$

in which cells of $L(< \sigma)$ (defined as $L(< \sigma) = \bigcup_{\tau \prec \sigma} L(\tau)$) are sent to 0 and cells of $L(\sigma)$ are sent to cells of $P(n - \#\sigma, \mathcal{W}(\sigma), w)$, up to sign. From the above discussion, one sees that $p_\sigma \circ i_\sigma = \mathrm{id}$. Together p_σ and i_σ give a splitting

(4.4) $$H_*(L(\leq \sigma)) \cong H_*(L(< \sigma)) \oplus H_{*-\#\sigma}\big(P(n - \#\sigma, \mathcal{W}(\sigma), w)\big).$$

This implies Theorem 4.1, once we produce the map spin_σ.

Finally, from the configuration point of view, the map $\mathrm{spin}_{\sigma,w}$ associates a configuration in the weighted no-$(w+1)$-equal space to a torus of configurations obtained by replacing points of weight r by wheels of size r and spinning those wheels. This also describes the action of i_σ on cycles. For example, when

$$\sigma = 3\ 7\ 5\ 9\ 8\ 10\ 13\ 4\ 15\ 17\ 14\ 18\ 20\ 21\ 12\ 2\ 22\ 16\ 23\ 6\ 11\ 1\ 24\ 19,$$

i_σ sends the 3–cycle in Figure 7 to the 14–cycle in Figure 4. For points in the core, the disks in the wheel are in the "standard" vertically ordered position.

Remark 4.5 Instead of using the splitting, one can prove Theorem 4.1 directly by constructing a discrete gradient vector field and applying Lemma 3.1. Put a total ordering $\prec$ on the cells of $\mathrm{cell}(n, w)$ as follows:

(i) If $\sigma(g) \prec \sigma(f)$, then $g \prec f$.

(ii) If $\sigma(g) = \sigma(f)$, then use the previously defined ordering on the set of cells of $P(n - \#\sigma, \mathcal{W}(\sigma), w)$. This ordering is based on an ordering on the wheels of σ; for this we order first by number of elements, then by axle.

This induces a discrete gradient vector field on cell(n, w) which restricts to that on $P(n - \#\sigma, \mathcal{W}(\sigma), w)$ on each layer. The images of the bases for $H_*\big(P(n - \#\sigma, \mathcal{W}(\sigma), w)\big)$ under i_σ for all σ give us a set of cycles to which we can apply Lemma 3.1.

For later reference, we describe the set of critical cells of this discrete gradient vector field in a self-contained way. Order wheels in a layer first according to their number of elements and then according to their largest element (*axle*). A block is a *unicycle* if it consists of a single wheel, that is, if its largest element comes first. We assign some blocks to leader–follower pairs by walking left to right: a block is a *follower* if it follows a unicycle (its *leader*) which is not itself a follower and whose wheel is smaller than any of the follower's wheels. A cell of cell(n, w) is critical if every block is either

(i)   a unicycle which is not a follower, or

(ii) a follower such that it and its leader have at least $w + 1$ elements in total.

4.2 Maps between permutohedra

To complete the proof of Theorem 4.1, we must still describe the map

$$\mathrm{spin}_\sigma \colon T^{\#\sigma} \times P(n - \#\sigma) \to \mathrm{cell}(n).$$

Informally, points in $P(n - \#\sigma)$ encode positions of the wheels of σ relative to each other, whereas points in the torus encode positions of disks inside the wheels, which can vary as in Figure 2. In particular, the image of spin_σ will be the union of the $2^{\#\sigma}$ top-dimensional cells obtained from σ by "spinning its wheels", that is, by applying permutations such as those depicted in Figure 9.

To make this precise, we start with the following lemma:

Lemma 4.6 *Let A be a finite set and $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ additional symbols not in A. Then there is a map*

$$\iota \colon [0, 1] \times P(A \cup \{\boldsymbol{a}\}) \to P(A \cup \{\boldsymbol{b}, \boldsymbol{c}\})$$

with the following properties:

(i)   *It is a homeomorphism and a cellular map. That is, it sends every k–face to a disk which is a union of k–faces.*
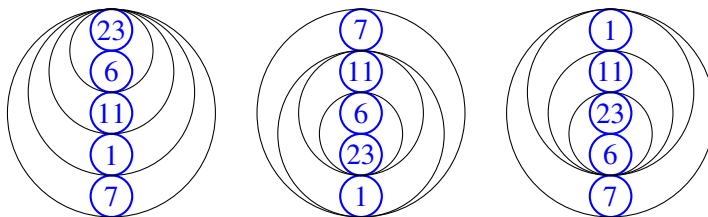

Figure 9: Some configurations of a wheel with five disks which give different orderings, the first of which is the "name" of the wheel.
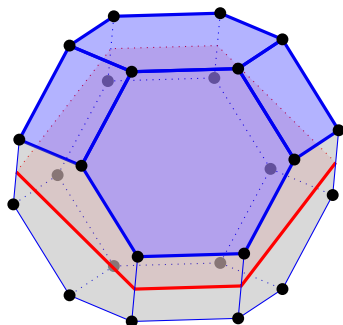
Figure 10: The permutohedron $P(4)$ and the homeomorphism $\iota\colon [0,1] \times P(3) \to P(4)$. The images of $\{\frac{1}{2}\} \times P(3)$ (red) and $\{1\} \times P(3)$ (blue) are highlighted.

(ii) *It is equivariant with respect to the $\mathbb{Z}/2\mathbb{Z}$–actions given by $t \mapsto -t$ on the domain and $\boldsymbol{b} \leftrightarrow \boldsymbol{c}$ on the codomain.*

(iii) *For each cell σ of $P(A \cup \{\boldsymbol{a}\})$, $\iota([0,1] \times \sigma)$ is the cell of $P(A \cup \{\boldsymbol{b}, \boldsymbol{c}\})$ with the same blocks except that $\boldsymbol{a}$ is replaced in its block by $\boldsymbol{b}$ and $\boldsymbol{c}$.*

(iv) *It takes $\{0\} \times P(A \cup \{\boldsymbol{a}\})$ to the union of those cells in which $\boldsymbol{b}$ and $\boldsymbol{c}$ are contained in separate blocks with $\boldsymbol{b}$ preceding $\boldsymbol{c}$. Similarly, it takes $\{1\} \times P(A \cup \{\boldsymbol{a}\})$ to the union of those cells in which $\boldsymbol{c}$ precedes $\boldsymbol{b}$.*

Before proving the lemma, we use it to construct spin_σ. First, notice that, for any ordering of $A \cup \{\boldsymbol{b}, \boldsymbol{c}\}$, the lemma gives us a well-defined map to the corresponding top cell of $\mathrm{cell}(A \cup \{\boldsymbol{b}, \boldsymbol{c}\})$. In particular, if we are instead given an ordering of $A \cup \{\boldsymbol{a}\}$ we have two choices: we can replace $\boldsymbol{a}$ by $\boldsymbol{b}\,\boldsymbol{c}$ or by $\boldsymbol{c}\,\boldsymbol{b}$. Moreover, these choices coincide on $\{0,1\} \times P(A \cup \{\boldsymbol{a}\})$. Identifying the two subspaces gives a map

$$S^1 \times P(A \cup \{\boldsymbol{a}\}) \to \mathrm{cell}(A \cup \{\boldsymbol{b}, \boldsymbol{c}\}).$$

Since top cells of $\mathrm{cell}(A \cup \{\boldsymbol{a}\})$ correspond to orderings of $A \cup \{\boldsymbol{a}\}$, this gives us a map

$$\mathrm{spin}\colon S^1 \times \mathrm{cell}(A \cup \{\boldsymbol{a}\}) \to \mathrm{cell}(A \cup \{\boldsymbol{b}, \boldsymbol{c}\}).$$

The equivariance of ι means that spin is well defined.

We build spin_σ by iterating the map spin. Start by thinking of $P(n-\#\sigma)$ as a top cell of $\mathrm{cell}(\{\text{wheels of }\sigma\})$. At each step, we replace a wheel with k elements by a singleton and a wheel with $k-1$ elements, applying spin to this splitting.

From this description, it is evident that

$$\mathrm{spin}_\sigma\big(P(n-\#\sigma, \mathcal{W}(\sigma), w)\big) = \mathrm{spin}_\sigma\big(P(n)\big) \cap \mathrm{cell}(n, w).$$

Proof of Lemma 4.6 In the proof, we will use a slightly different notation than in the statement of the lemma: we will replace the finite set A by $\{1, \ldots, n-2\}$ and the symbols $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ by $n-1$, $n-1$ and n, respectively. This lets us write coordinates in $\mathbb{R}^n$ more easily.

The permutohedron $P(n)$ is a *zonotope*, that is, the Minkowski sum of a set of line segments. We refer to [Ziegler 1995, Section 7.3] for known facts about zonotopes.

The standard permutohedron $P(n)$ is the zonotope in $\mathbb{R}^n$ which is the Minkowski sum of the $\binom{n}{2}$ segments connecting every pair of standard unit basis vectors. That is,

$$P(n) = \left\{ \sum_{1 \leq i < j \leq n} a_{ij} e_i + (1 - a_{ij}) e_j \;\middle|\; 0 \leq a_{ij} \leq 1 \right\}.$$

Evidently, choosing a different set of n linearly independent vectors in any Euclidean space gives a linearly equivalent polytope. Less obviously, the combinatorial structure of a zonotope only depends on the oriented matroid associated to the set of line segments. In particular, changing the lengths of the line segments without changing their direction does not change the combinatorial structure.

We use this fact to prove the following:

Lemma 4.7 *The polytope $Z(n) = P(n) \cap \{x_{n-1} = x_n\}$ is combinatorially equivalent to $P(n-1)$.*

Proof We first show that $Z(n)$ can also be written as

$$(4.8) \qquad \left\{ \sum_{1 \leq i < j \leq n} a_{ij} e_i + (1 - a_{ij}) e_j \;\middle|\; 0 \leq a_{ij} \leq 1, \, a_{i(n-1)} = a_{in}, \, a_{(n-1)n} = \tfrac{1}{2} \right\} \subset P(n).$$

Indeed, suppose that $z \in Z(n)$, so that

$$z = \sum_{1 \leq i < j \leq n} b_{ij} e_i + (1 - b_{ij}) e_j \in P(n)$$

and $z_{n-1} = z_n$. If we switch e_{n-1} and e_n in this formula, we get z back. We can get an expression as in (4.8) by averaging these two expressions for z, getting

$$a_{i(n-1)} = a_{in} = \tfrac{1}{2}(b_{i(n-1)} + b_{in}), \quad a_{(n-1)n} = \tfrac{1}{2}.$$

From (4.8), we see that $Z(n)$ is the zonotope generated by the segments

$$e_i \leftrightarrow e_j \quad \text{for } 1 \leq i < j \leq n-2 \qquad \text{and} \qquad e_{n-1} + e_n \leftrightarrow 2e_i \quad \text{for } 1 \leq i \leq n-2.$$

This zonotope is combinatorially equivalent to one in which segments $e_{n-1} + e_n \leftrightarrow 2e_i$ are replaced by $\tfrac{1}{2}(e_{n-1} + e_n) \leftrightarrow e_i$. This in turn is linearly equivalent to $P(n-1)$. □

The lemma gives us an injective map $\iota_{1/2} \colon P(n-1) \to P(n)$ which is a homeomorphism onto its image; it remains to extend it to $\iota \colon [0, 1] \times P(n-1) \to P(n)$. We do this by letting

$$\iota(t, x) = \iota_{1/2}(x) + R(x)(2t - 1)(e_n - e_{n-1}),$$

where $R(x)$ is the maximum value for which the image still lies in $P(n)$. Property (ii) of the lemma follows immediately from this definition.

To show that ι is the desired map, we must show it is a homeomorphism. To prove injectivity, it suffices to show that $R(x)$ is always positive. But in fact $R(x) \geq \tfrac{1}{2}$, since we can always vary $a_{(n-1)n}$. Surjectivity

follows from the fact that $P(n)$ is convex and symmetric about the hyperplane $Z(n)$. Finally, ι can only be discontinuous if $R \circ \iota_{1/2}^{-1} \colon Z(n) \to \mathbb{R}$ is discontinuous, but convexity of $P(n)$ implies that this is a convex function, and therefore continuous. Continuity of the inverse likewise follows from the fact that the reciprocal of this function is continuous.

Next we show that ι is cellular. Given a face σ of $P(n-1)$, we know that $\iota_{1/2}(\sigma)$ is a face of $Z(n)$ cut out by a half-space of the hyperplane $\{x_{n-1} = x_n\}$. Extending this half-space orthogonally to $\mathbb{R}^n$, we obtain the half-space that cuts out the face $\iota([0,1] \times \sigma)$ of $P(n)$. It follows that $\iota(\{0\} \times \sigma)$ and $\iota(\{1\} \times \sigma)$ are also unions of faces of $P(n)$ of the appropriate dimension: specifically, each facet of $\iota([0,1] \times \sigma)$ which is not the image of $[0,1] \times \tau$ for some facet τ of σ is contained in one of those two, depending on whether $x_{n-1} - x_n$ is positive or negative for points in that facet.

To show properties (iii) and (iv), we recall the relationship between symbols of cells and the zonotope structure. Namely, points in a face corresponding to a given symbol are those which can be expressed as

$$\sum_{1 \le i < j \le n} a_{ij} e_i + (1 - a_{ij}) e_j,$$

where a_{ij} is 0 if j is in a later block than i and 1 if i is in a later block than j. When i and j are in the same block, a_{ij} can be anything in $[0,1]$.

From the proof of Lemma 4.7, it follows that $\iota_{1/2}$ maps a face with a given symbol into a face with the same symbol with n added to the same block as $n-1$. Since ι is cellular, this shows (iii). Property (iv) then also follows from our argument that ι is cellular. □

4.3 Proofs of main theorems

Proof of Theorem B Combining the arguments in Section 3 and the proof of Theorem 4.1, we find a basis for $H_*(\mathrm{cell}(n,w))$: for each permutation $\sigma \in S_n$ and each basic cycle z of $H_j\big(P(n-\#\sigma, \mathcal{W}(\sigma), w)\big)$, it contains the cycle $i_\sigma(z)$. The exact set of cycles we get depends on the correspondence between wheels of σ and elements of $[n - \#\sigma]$; to make this concrete, we order the wheels first by size and then by axle.

Geometrically, the correspondence $z \mapsto i_\sigma(z)$ matches

- points moving in $\mathbb{R}$ to wheels moving in $\mathbb{R} \times [0, w]$;
- points moving through each other to wheels moving over and above each other, with the wheel with the smaller axle on top.

We now verify the numbered conditions of Theorem B:

- Condition (i) follows from the way we split σ into wheels.
- Condition (ii) follows from the definition of the layer $L(\sigma)$.
- Conditions (iii) and (iv) follow from Lemma 3.2 and the ordering of the wheels. □

Proof of Theorem A Every basic cycle is a concatenation product of wheels and filters. It is enough to show that there are finitely many (unlabeled) types of wheels and filters in $\text{config}(-, w)$, and that these have at most $\frac{3}{2}w$ disks and generate cycles of dimension at most $\frac{3}{2}w - 2$.

A wheel consists of at most w disks, and a wheel with k disks generates a $(k-1)$–cycle. Filters with at least three wheels have at most $\frac{3}{2}w$ disks, and a filter with k disks generates a $(k-2)$–cycle. Filters with two wheels b_1 and b_2 can be written as $b_2|b_1 - b_1|b_2$, and do not need to be counted separately. Therefore, $H_*(\text{config}(-, w))$ is spanned by concatenation products of cycles in $H_{\leq 3/2w-2}\big(\text{config}\big(\leq \frac{3}{2}w, w\big)\big)$.

Every wheel with the same number of disks has the same shape, and there are finitely many shapes of filters of width w. Therefore, $H_*(\text{config}(-, w))$ is finitely generated as a twisted algebra. $\qquad\square$

5 Betti number growth function

Alpert et al. [2021] examine the growth of the dimension of $H_j(\text{config}(n, w))$ for fixed j and w and varying n, and show that it is asymptotically polynomial times exponential. Having computed a basis for homology, we can now answer the question, is the dimension exactly equal to a polynomial times an exponential? The answer is that it is a sum of such functions, each with a different base of the exponential. The polynomials are integer-valued, and the theory of finite difference calculus states that every integer-valued polynomial in n is an integer linear combination of binomial coefficient functions $\binom{n}{a}$ for various a. Thus, in this section we prove the following theorem:

Theorem 5.1 *For fixed j and w, as a function of n the dimension of $H_j(\text{config}(n, w))$ is an integer linear combination of terms of the form $\binom{n}{a}b^{n-a}$, where a and b are nonnegative integers.*

Note that the case $b = 0$ is permitted, and represents the function that is equal to 1 when $n = a$, and 0 otherwise.

Our proof counts the critical cells from Remark 4.5, and decomposes them into concatenation products of factors that are easier to count. To describe how the counts transform under concatenation product, we introduce the following terminology. A *cell family* F consists of, for each n, a set F_n of cells from cell(n). Its *counting function* $f(n)$ is the number of cells in F_n. The concatenation product $F \mid G$ of cell families F and G is the cell family consisting of all concatenation products of a cell in F with a cell in G, taken with all possible disk labelings that preserve the order within each factor.[2] The following lemma shows that, to prove our theorem, it suffices to consider separately the various independent factors of a concatenation product:

Lemma 5.2 *Let F and G be two cell families, such that their counting functions $f(n)$ and $g(n)$ are integer linear combinations of terms of the form $\binom{n}{a}b^{n-a}$, where a and b are nonnegative integers. Suppose that, for every element of their concatenation product $F \mid G$, there is only one way to write it as*

---

[2]This is closely related to the Day convolution defined in Section 2.3, except that F and G need not be FB–subsets of the FB–set of cells of cell$(-)$, and in our application will not be S_n–equivariant.

the concatenation of an element of F and an element of G. Then the counting function of $F \mid G$ is also an integer linear combination of terms of the form $\binom{n}{a} b^{n-a}$.

Proof Let $f \mid g$ denote the counting function of $F \mid G$. Then we have

$$(f \mid g)(n) = \sum_{i+j=n} \binom{i+j}{i} f(i) g(j),$$

which we refer to as the *labeled convolution* of f and g. Exponential generating functions are convenient for dealing with these labeled convolutions; given the exponential generating functions $\sum_i f(i) x^i / i!$ for f and $\sum_j g(j) x^j / j!$ for g, their product gives the exponential generating function $\sum_n (f \mid g)(n) x^n / n!$ for $f \mid g$.

We can compute the labeled convolution of two terms $\binom{n}{a_1} b_1^{n-a_1}$ and $\binom{n}{a_2} b_2^{n-a_2}$ by converting them to exponential generating functions, taking the product, and then converting the product back. The exponential generating function of $\binom{n}{a} b^{n-a}$ is

$$\sum_n \binom{n}{a} b^{n-a} \frac{x^n}{n!} = \sum_n \frac{x^a}{a!} \frac{(bx)^{n-a}}{(n-a)!} = \frac{x^a}{a!} e^{bx},$$

so the product of exponential generating functions of $\binom{n}{a_1} b_1^{n-a_1}$ and $\binom{n}{a_2} b_2^{n-a_2}$ is

$$\frac{x^{a_1+a_2}}{a_1! a_2!} e^{(b_1+b_2)x} = \binom{a_1+a_2}{a_1} \frac{x^{a_1+a_2}}{(a_1+a_2)!} e^{(b_1+b_2)x},$$

and so the labeled convolution of $\binom{n}{a_1} b_1^{n-a_1}$ and $\binom{n}{a_2} b_2^{n-a_2}$ is

$$\binom{a_1+a_2}{a_1} \binom{n}{a_1+a_2} (b_1+b_2)^{n-(a_1+a_2)}.$$

(This identity can also be verified by constructing sets counted by each of the counting functions.) $\qquad \square$

Using this lemma, we are ready to prove Theorem 5.1.

Proof of Theorem 5.1 Given j and w, the critical cells from Remark 4.5 form a cell family, and we want to find the counting function of this cell family. To do this, it helps to count the ways to distribute the singletons separately from counting the ways to arrange the larger wheels. We define a *skyline* to be a critical cell for which the only singleton blocks are leaders of a filter. Given any critical cell, we can find its associated skyline by deleting all the nonleader singletons and then shifting the disk labels down so they are consecutive. For fixed j and w, there are only finitely many possible skylines among the critical cells. Thus, it suffices to count the cell family consisting of all critical cells with a given skyline.

For each skyline, its cell family is the concatenation product of other cell families. We define a *tadpole* to be a critical cell with exactly one filter, which is at the far right, and we define a *tail* to be a critical cell

with no filters. Every critical cell can be written uniquely as a concatenation of some number of tadpoles and possibly one tail on the right. Thus, it suffices to count the cell family consisting of all tadpoles with a given skyline, and the cell family consisting of all tails with a given skyline.

A tadpole with a singleton as the leader of its filter may have other singletons to the left, but a tadpole with a larger wheel as the leader of its filter may not have any singletons. Thus, given a tadpole skyline, if the leader of its filter is not a singleton, the counting function of its cell family returns 1 when the input is the number of disks in the skyline, and 0 otherwise. If the leader of its filter is a singleton, then, for any tadpole with that skyline, the leader of the filter must be disk 1. If k is the total number of disks in the skyline, then the counting function of the tadpole skyline family is $\binom{n-1}{k-1}$, representing the number of ways to choose which disk labels appear in the skyline. Given a tail skyline, if k is the total number of disks in the skyline (possibly 0), the counting function of its cell family is $\binom{n}{k}$.

Thus, for each skyline with a given j and w, we can compute the counting function of its cell family by taking the labeled convolution of the counting functions of its tadpole skyline and tail skyline factors. Because each factor has the desired form, so does the labeled convolution, by Lemma 5.2. The counting function of the family of all critical cells is the sum of the counting functions of all the skylines, and so it also has the desired form. $\qquad\square$

6 Cohomology ring

We now give a basis for $H^j(\mathrm{config}(n, w))$ and describe the cup product structure in terms of that basis. We use a similar strategy to that used in [Alpert et al. 2021] to find a lower bound for the dimensions of homology groups. The main tool is Poincaré–Lefschetz duality: for a compact $2n$–manifold with boundary $(M, \partial M)$,

$$H^j(M) \cong H_{2n-j}(M, \partial M).$$

Moreover, when homology classes are realized by submanifolds (as they will be in our case), then:

(i) The pairing between classes in $H^j(M)$ and $H_j(M)$ is given by the transverse signed intersection number between classes in $H_{2n-j}(M, \partial M)$ and $H_j(M)$.

(ii) The cup product between classes in $H^i(M)$ and $H^j(M)$ is given by the transverse intersection map
$$\pitchfork\colon H_{2n-i}(M, \partial M) \otimes H_{2n-j}(M, \partial M) \to H_{2n-i-j}(M, \partial M).$$

While $\mathrm{config}(n, w)$ as previously described is not a compact $2n$–manifold with boundary, we can define a homotopy equivalent compact manifold with boundary:

Definition Let $M(n, w) \subset \mathbb{R}^{2n}$ be the configuration space of open disks of radius 1 in a strip of any finite length $N > n$ and width $w + \varepsilon$ for any $0 < \varepsilon < 1$.

The boundary consists of those configurations in which a disk touches either another disk or the boundary of the strip. It has corners, but every point of the boundary has a neighborhood homeomorphic to a half-space. (Without the addition of ε, a boundary configuration with w vertically aligned disks would not have a neighborhood homeomorphic to a half-space.)

Alternatively, in an open manifold without boundary (such as the space of configurations of n points in $\mathbb{R} \times (0, 1)$ of which no more than w are vertically aligned, as used in Theorem 2.1), Poincaré duality gives an isomorphism $H^j(M) \cong H^{\mathrm{BM}}_{2n-j}(M)$. Here H^{BM}_* indicates *Borel–Moore homology*, the homology of the complex of locally finite chains; this is isomorphic to the homology of a compactification relative to the added points.

6.1 Basis for cohomology

We will describe a basis for $H_{2n-j}(M(n, w), \partial M(n, w))$ by associating basis elements to critical j–cells of cell(n, w), as described in Remark 4.5. Given a critical cell with symbol f, we define the submanifold $V(f) \subset M(n, w)$ as the set of configurations such that:

   (i) The disks in each block of σ are lined up vertically in order.
   (ii) If a block b_1 comes before b_2 in f, they have at least $w + 1$ elements combined, and one of them is a follower, then the column of disks labeled by elements of b_1 is to the left of that labeled by elements of b_2.

To see that this is a submanifold, and that moreover $\partial V(f) = V(f) \cap \partial M(n, w)$, notice that two columns of disks which have at least $w + 1$ elements combined cannot move past each other while still satisfying condition (i).

To show that this is a basis, we describe the intersection pairing between these submanifolds and our generators of $H_j(\mathrm{config}(n, w))$.

Lemma 6.1 *Let f and g be symbols of critical cells of* cell(n, w)*. Write $Z(g)$ for the basic cycle corresponding to g. Then*:

   (a) $V(g) \pitchfork Z(g) = \pm 1$.
   (b) *If $V(g) \pitchfork Z(f) \neq 0$, then $g \preceq f$ according to the ordering in* Remark 4.5.

From the lemma, we see that, under the ordering of the critical cells by $\prec$, the intersection pairing is described by a triangular matrix. Therefore, the $V(g)$ form a basis for $H_{2n-j}(M(n, w), \partial M(n, w))$.

Proof We use the embedding of Theorem 2.1 to associate $V(g)$ to a cellular j–cocycle $\nu(g)$ in cell(n, w). The value of $\nu(g)$ on a j–cell is given by the signed intersection number of the embedded cell with $V(g)$.

Suppose that a symbol h is in the support of $\nu(g)$. By comparing condition (i) of the definition of $V(g)$ to the construction of the embedding in Theorem 2.1, we see that f must have the same set of blocks as g, although possibly in a different order; in particular, h and g are in the same layer. Moreover, the barycenter of h is the unique point of intersection. Condition (ii) restricts the possible orderings.

Now suppose that $h \neq g$. We look at the first block, say h_i and g_i, where they differ. This means that h_i is a block which occurs later in g_i, say g_j. The structure of critical cells gives us the following possibilities:

- The block g_i is a follower in g, and g_j is not. Since g_i is a follower, it and g_{i-1} have at least $w + 1$ elements in total. On the other hand, condition (ii) implies that g_i and g_j have at most w elements in total, so g_{i-1} has more elements than g_j. Since g_j is not a follower, it must be a unicycle. Then $g_{i-1} \prec g_j = h_i$ in the ordering on wheels, implying that h_i is not a follower in h. Therefore, $g \prec h$.

- There are no followers between g_i and g_j, inclusive. Then g_j and g_i are both unicycles and $g_j \prec g_i$ in the ordering on wheels. Therefore, h_i is not a follower and $h_i \prec g_i$. Because the ordering on cells uses the reverse ordering on wheels, $g \prec h$.

- There is at least one follower between g_i and g_j, and it is not g_i. But g_j can't be a follower, because if it were, $V(g)$ could not contain configurations with g_j to the left of its leader. Then g_j and the first follower after g_i appear in opposite orders in g and h, so g_j must have fewer elements than the first leader after g_i. Since g_j is not a follower, it is a unicycle, and the ordering implies that it must also have fewer elements than g_i. Thus, $h_i \prec g_i$, and therefore $g \prec h$.

In other words, we always have that $g \preceq h$. On the other hand, we know that, if a cell h is in the support of $Z(f)$, then $h \preceq f$. Therefore, $g \preceq f$. This proves (b).

From this we also know that g is the unique cell which is in the support of both $\nu(g)$ and $Z(g)$. Since the coefficient in both cases is ± 1, this proves (a). □

6.2 Cup product structure

The cup product structure of $H^*(\mathrm{config}(n, w))$ is complicated, with many indecomposables as well as many nontrivial products. We start with some simple observations. First, the cohomology algebra of $H^*(\mathrm{config}(n))$ is well understood: it is generated by one-dimensional classes. A good description is given in [Sinha 2013]. Secondly, the pullback of $H^*(\mathrm{config}(n))$ to $H^*(\mathrm{config}(n, w))$ along the inclusion map is a subalgebra and contains all classes of degree less than $w - 1$. That means that all classes in $H^{w-1}(\mathrm{config}(n, w))$ which are not pullbacks from $H^{w-1}(\mathrm{config}(n))$ are indecomposable. These include the basis elements corresponding to critical cells which have one leader–follower pair with $w + 1$ total elements and in which all other blocks are singletons.

In higher degrees, the story is more complicated. Recall that our basic cocycles are carved out using three kinds of relations: coincidence between the x–coordinates of two disks, vertical ordering of elements within a block, and horizontal ordering between blocks. When we take a cup product, the coincidences from both factors accumulate; see Table 1 for some examples. Likewise, when two blocks are ordered in a product, the ordering must be "inherited" from one of the factors. If many pairs of blocks are ordered, this may force the cohomology class to be indecomposable.

$$\nu(2\,1\,|\,6\,|\,5\,|\,4\,|\,3) \cup \nu(3\,2\,|\,6\,|\,5\,|\,4\,|\,1) = \nu(3\,2\,1\,|\,6\,|\,5\,|\,4)$$

$$\nu(3\,2\,|\,6\,|\,5\,|\,4\,|\,1) \cup \nu(3\,1\,|\,6\,|\,5\,|\,4\,|\,2) = \nu(3\,2\,1\,|\,6\,|\,5\,|\,4) + \nu(3\,1\,2\,|\,6\,|\,5\,|\,4)$$

$$\nu(3\,2\,1\,|\,6\,|\,5\,|\,4) \cup \nu(5\,4\,|\,6\,|\,3\,|\,2\,|\,1) = \nu(3\,2\,1\,|\,5\,4\,|\,6)$$

$$\nu(5\,4\,|\,6\,2\,3\,|\,1) \cup \nu(4\,1\,|\,6\,|\,5\,|\,3\,|\,2) = \nu(5\,4\,1\,|\,6\,2\,3).$$

Table 1: Some examples of nontrivial cup products in $\mathrm{config}(6,4)$. These are easy to deduce using intersections of dual cycles.

The last example in Table 1 illustrates an important class of decomposable cohomology classes: those associated to critical cells consisting of only two blocks, where the total number of elements is at least $w + 2$. In other words, a filter with more than $w + 1$ elements always pairs with a decomposable cohomology class.

We also describe an important set of cases in which cup products are zero.

Proposition 6.2 *If there is a pair of labels i and j which are contained in the same block in both f and g, then $V(f) \pitchfork V(g) = \varnothing$.*

Proof If the two labels are in opposite orders in the two blocks, then the intersection of the two cycles is empty. If they are in the same order, then the intersection may be nonempty, but we can move it off itself by moving every point of $V(f)$ slightly in the x_i–direction (say, move the point p by a distance of $\varepsilon d(p, \partial M(n, w))$ for some $\varepsilon > 0$ small enough). After this operation, $V(f) \cap V(g)$ lies in the boundary of $M(n, w)$. $\square$

In all other cases, we can compute cup products by looking at the intersections of the associated submanifolds.

Proposition 6.3 *If no two blocks in f and g have two labels in common, then $V(f)$ and $V(g)$ intersect transversely.*

Proof Locally, $V(f)$ and $V(g)$ are linear subspaces of $\mathrm{config}(n)$, cut out by the linear equations constraining the x–coordinates in each block to coincide. Thus, the dimension of $V(f)$ is n (for the y–coordinates) plus the number of blocks in f (for the x–coordinates), and similarly for $V(g)$. In the intersection, we imagine starting with the constraints for $V(f)$ and including the constraints for $V(g)$ one block at a time. Each block of size k in g merges k different blocks in f into one, so the net change in the number of blocks is $1 - k$. In total, the number of blocks in $V(f) \cap V(g)$ is $\#\mathrm{blocks}(f) + \#\mathrm{blocks}(g) - n$, so the local codimension of $V(f) \cap V(g)$ is $2n - \#\mathrm{blocks}(f) - \#\mathrm{blocks}(g)$, which equals $\mathrm{codim}\, V(f) + \mathrm{codim}\, V(g)$. $\square$

Finally, we show that there are many indecomposable elements in $H^*(\mathrm{config}(n, w))$, but that they do not occur in the very highest degrees.

Theorem 6.4 *The ring $H^*(\text{config}(n, w))$ has indecomposables*

(a) *only in degree 1 when $w = 2$;*

(b) *in every degree between 1 and $\lfloor \frac{1}{2}n \rfloor$ and no others when $w = 3$;*

(c) *in degree 1 and in every degree between $w - 1$ and $\lfloor \frac{1}{2}(n + w - 3) \rfloor$ and no degree greater than $n - \lceil n/(w-1) \rceil$ when $w \geq 4$.*

When $w \geq 4$, indecomposables also seem to occur in most degrees below $n - \lceil n/(w-1) \rceil$, but perhaps not all.

Proof We first show that every class of degree greater than $n - \lceil n/(w-1) \rceil$ is decomposable. The proof does not depend on w. Every cell with fewer than $\lceil n/(w-1) \rceil$ blocks has a block with w elements, and therefore $V(f) \subset M(n, w)$ satisfies an equation of the form $x_{i_1} = \cdots = x_{i_w}$. Therefore, it is enough to show:

Lemma 6.5 *Suppose that $V \subset M(n, w)$ is a connected compact submanifold of dimension at most $2n - w$, satisfying $\partial V \subset \partial M(n, w)$, which is cut out by relations of the form $x_i = x_j$, $y_i < y_j$ and $x_i < x_j$. Suppose furthermore that V satisfies $x_{i_1} = \cdots = x_{i_w}$ for some set of indices $i_1, \ldots, i_w$. Then V is the transverse intersection of two proper compact submanifolds satisfying $\partial V \subset \partial M(n, w)$.*

Proof Define W_1 to be the connected component of $\{x_{i_1} = \cdots = x_{i_w}\}$ which contains V; this exists since V is connected. The defining relations of W_1 are:

- $x_{i_k} = x_{i_l}$ for each $k \neq l$.
- $y_{i_k} < y_{i_l}$ or $y_{i_k} > y_{i_l}$ for each $k \neq l$.
- $x_j < x_{i_k}$ or $x_j > x_{i_k}$ whenever $j \neq i_k$ for any k.

Let W_2 be the submanifold cut out by all defining relations of V which involve pairs of points constrained to be to the same side of $x_{i_1}, \ldots, x_{i_w}$. Then $V = W_1 \cap W_2$, since every necessary relation defining V is a defining relation of either W_1 or W_2. Moreover, by counting the number of defining equalities, we immediately see that the intersection is transverse. □

Applying this to each connected component of $V(f)$, we get a decomposition of the corresponding cohomology class as a sum of cup products.

We now build indecomposable cocycles when $w \geq 4$; the proof for $w = 3$ will be similar, but not identical. We will show that basic cocycles corresponding to critical cells of certain shapes are indecomposable. Specifically, we consider a critical cell f which starts with some number r of blocks with two elements, followed by one block (a follower) with $w - 1$ elements, and where the remaining blocks are singletons. The degree of such an f is $r + w - 2$, and r can be any number between 1 and $\frac{1}{2}(n - w + 1)$, giving us all degrees between $w - 1$ and $\lfloor \frac{1}{2}(n + w - 3) \rfloor$. It is clear that there are critical cells of any such shape; in particular, we select f to be the cell with all entries in order from greatest to least.

We will show that the corresponding cohomology class $\nu(f)$ is indecomposable by induction on r. We will do this by constructing a cycle with which $\nu(f)$ pairs nontrivially, but any decomposable class pairs trivially.

To construct this cycle, first let R be the set of blocks of f which are left of the follower. For any subset $S \subseteq R$, there is a critical cell f_S in which the blocks in S are moved to the right of the follower, and otherwise the ordering on blocks is the same. Let Z_S be the concatenation product of the toroidal cycles obtained by interpreting each block as a wheel and spinning it. This cycle is represented by a map $g_S \colon T^{r+w-2} \to \mathrm{config}(n, w)$. The Z_S represent linearly independent homology classes, since they generate a 2^r–dimensional subspace of $H_{r+w-2}(\mathrm{config}(n, w))$: for every $S \subseteq R$, the basic cycle $Z(f_S)$ is a linear combination of them, given by Z_S itself if $S = R$, or otherwise the difference between Z_S and $Z_{S'}$, where S' is the union of S with the greatest block not in S. We will show that, for every decomposable class ν,

$$\text{(6.6)} \qquad \sum_{S \subseteq R} (-1)^{|S|} \langle \nu, Z_S \rangle = 0.$$

In particular, ν cannot pair nontrivially with exactly one of the cycles Z_S. On the other hand, $\nu(f)$ pairs nontrivially with Z_S if and only if $S = \varnothing$.

It is enough to show this equation holds for every pairwise cup product, $\nu = \alpha \cup \beta$. We study the pullbacks of such a pairwise cup product along each g_S. Suppose that α is of degree p, and write

$$T^{r+w-2} = \prod_{i \in R \sqcup R'} S_i^1,$$

where R' is the set of degrees of freedom of the $(w-1)$–element wheel. To understand $g_S^* \alpha$, it is enough to understand how it pairs with $T^P = \prod_{i \in P} S_i^1$ for each p–element set $P \subseteq R \sqcup R'$. Then

$$\text{(6.7)} \qquad \langle \nu, Z_S \rangle = \sum_{P \sqcup Q = R \sqcup R'} \langle g_S^* \alpha, [T^P] \rangle \cdot \langle g_S^* \beta, [T^Q] \rangle.$$

We now show that these decompositions are not independent for different S.

Lemma 6.8 *If $P \cap R' \neq R'$, then $\langle g_S^* \alpha, [T^P] \rangle$ is the same for every S.*

Proof It suffices to show that the pushforward cycles $(g_S)_* [T^P]$ are homologous regardless of S. In fact, the corresponding maps $T^P \to \mathrm{config}(n, w)$ are homotopic. We can compress the allowed movements of the $(w-1)$–element wheel into a subset of the strip of width $w - 2$, letting wheels of width 2 pass by. Clearly, wheels of width 2 can also pass by each other since we are assuming $w \geq 4$. Using this set of motions, we can construct a homotopy between any two such maps. □

Lemma 6.9 *If $P \cap R' = R'$, then $\langle g_S^* \alpha, [T^P] \rangle = \langle g_{S'}^* \alpha, [T^P] \rangle$ whenever $(S \bigtriangleup S') \cap P = \varnothing$ (where $\bigtriangleup$ indicates symmetric difference).*

Proof Again, it suffices to show that $(g_S)_*[T^P]$ and $(g_{S'})_*[T^P]$ are homologous, and again the corresponding maps $T^P \to \text{config}(n, w)$ are homotopic. We can build the homotopy by moving the individual disks of blocks not in P around the $(w-1)$–element set. □

Together, the lemmas imply that, for any pair of nonempty complementary sets $P, Q \subset R \sqcup R'$, the quantity

$$\langle g_S^* \alpha, [T^P] \rangle \cdot \langle g_S^* \beta, [T^Q] \rangle$$

is independent of whether $i \in S$ for at least one $i \in R$. In particular,

$$\sum_{S \subseteq R} (-1)^{|S|} \langle g_S^* \alpha, [T^P] \rangle \cdot \langle g_S^* \beta, [T^Q] \rangle = 0.$$

From here we get (6.6) by summing over all pairs P and Q and using (6.7).

Finally, we deal with the case $w = 3$. In this case, we consider critical cells composed of r two-element blocks (with the bigger element first) followed by $n - 2r$ singletons for some $1 \leq r \leq \lfloor \frac{1}{2} n \rfloor$. Such a cell is critical if and only if the singletons are in reverse order. In particular, every permutation σ of the set of two-element blocks gives a critical cell f_σ. For each $\sigma \in S_r$, let Z_σ be an r–cycle, represented by a map $g_\sigma : T^r \to \text{config}(n, w)$, obtained by arranging the blocks in the order dictated by σ and spinning each of them. These cycles are linearly independent in homology since the basic cycles $Z(f_\sigma)$ are all linear combinations of them.

We now define a function $\mu : S_r \to \mathbb{Z}$. Consider the expression $[\cdots [[1, 2], 3], \cdots r]$. If σ cannot be obtained from this by commuting some of the brackets (equivalently, by spinning the wheel $1 \, 2 \cdots r$, as in Figure 9), then $\mu(\sigma) = 0$. If it can, then $\mu(\sigma) = (-1)^c$, where c is the number of commutations required. We will show that, for every decomposable class v,

(6.10) $$\sum_{\sigma \in S_r} \mu(\sigma) \langle v, Z_\sigma \rangle = 0.$$

In particular, v cannot pair nontrivially with exactly one of the Z_σ, and therefore there is an indecomposable class of degree r.

It is enough to show the equation for every pairwise cup product, $v = \alpha \cup \beta$. Write $T^r = \prod_{i=1}^r S_i^1$; for every $P \subset \{1, \ldots, r\}$, write $T^P = \prod_{i \in P} S_i^1$. Then

(6.11) $$\langle v, Z_\sigma \rangle = \sum_{P \sqcup Q = \{1, \ldots, r\}} \langle g_\sigma^* \alpha, [T^P] \rangle \cdot \langle g_\sigma^* \beta, [T^Q] \rangle.$$

Once again, these decompositions are not independent for different σ:

Lemma 6.12 *Whenever σ and τ impose the same ordering on elements of P,*

$$\langle g_\sigma^* \alpha, [T^P] \rangle = \langle g_\tau^* \alpha, [T^P] \rangle.$$

Proof To show that $(g_\sigma)_*[T^P]$ is homologous to $(g_\tau)_*[T^P]$, we construct a homotopy between the corresponding maps $T^P \to \text{config}(n, w)$. This involves moving around the individual disks of the blocks not in P to make sure they are in the right order. □

The lemma implies that, for any pair of nonempty complementary sets $P, Q \subset \{1, \dots, r\}$,

$$\langle g_\sigma^* \alpha, [T^P] \rangle \cdot \langle g_\sigma^* \beta, [T^Q] \rangle = \langle g_\tau^* \alpha, [T^P] \rangle \cdot \langle g_\tau^* \beta, [T^Q] \rangle$$

if τ and σ differ by commuting one of the brackets of $[\cdots[[1, 2], 3], \cdots r]$, specifically the innermost bracket in which the right side is in Q if $1 \in P$, or vice versa. In particular,

$$\sum_{\sigma \in S_r} \mu(\sigma) \langle g_\sigma^* \alpha, [T^P] \rangle \cdot \langle g_\sigma^* \beta, [T^Q] \rangle = 0.$$

From here we get (6.10) by summing over all pairs P and Q and using (6.11). □

7 Persistent homology

The majority of this paper has investigated the properties of $\text{config}(n, w)$ as we increase n and keep w fixed. But we can also look at what happens when w grows. Since $\text{config}(n, w)$ naturally injects into $\text{config}(n, w + 1)$, forming a filtration, the right framework for understanding this is *persistent homology*, which considers homology for all w at once, together with the maps induced by the inclusions. We give a short introduction to the machinery; for more details, see [Edelsbrunner and Harer 2008; Zomorodian and Carlsson 2005].

Specifically, we can regard $\bigoplus_w H_*(\text{config}(n, w))$ as a $\mathbb{Z}[t]$–module in which multiplication by t corresponds to applying the maps $H_*(\text{config}(n, w)) \to H_*(\text{config}(n, w + 1))$ induced by the inclusions $\text{config}(n, w) \hookrightarrow \text{config}(n, w + 1)$. We denote this $\mathbb{Z}[t]$–module by $\text{PH}_*(\text{config}(n, *))$. Similarly, $\bigoplus_w H^*(\text{config}(n, w))$ is a $\mathbb{Z}[t]$–module in which multiplication by t corresponds to applying the pullback maps $H^*(\text{config}(n, w)) \to H^*(\text{config}(n, w - 1))$, and we denote this $\mathbb{Z}[t]$–module by $\text{PH}^*(\text{config}(n, *))$.

A *cyclic* summand of $\text{PH}_*(\text{config}(n, *))$ or $\text{PH}^*(\text{config}(n, *))$ is generated by a single element that is nonzero for some interval of values of w. It is standard to refer to a cyclic summand as a *bar*, and to the endpoints of the corresponding interval as the values w of its *birth* and *death*. A decomposition of $\text{PH}_*(\text{config}(n, *))$ or $\text{PH}^*(\text{config}(n, *))$ into cyclic summands means selecting $\mathbb{Z}$–bases for the various values of w in a way that agrees with the maps between the values of w.

The fundamental theorem for modules over a PID guarantees that a persistence module with coefficients in a field will always decompose into cyclic summands. No such guarantee exists for integral persistence modules: for example, one could have a single class born at one time and later become divisible by 2, yielding a module isomorphic to the ideal $(2, t) \subset \mathbb{Z}[t]$.

Theorems 7.1 and 7.4, which we state and prove below, together prove Theorem C.

Theorem 7.1 *The homology basis elements described in Theorem B induce a decomposition of* $\mathrm{PH}_*(\mathrm{config}(n, *))$ *as the direct sum of cyclic* $\mathbb{Z}[t]$*–modules. The cohomology basis elements from Section 6.1 give a decomposition of* $\mathrm{PH}^*(\mathrm{config}(n, *))$ *as the direct sum of cyclic* $\mathbb{Z}[t]$*–modules.*

To prove the theorem, it suffices to show that, when a given basis element stops being in the basis, it also becomes zero in homology or cohomology. To verify this for homology, we show the corresponding statement for weighted no-$(w+1)$-equal spaces.

Theorem 7.2 *For the weighted no-$(w+1)$-equal spaces, the homology basis elements from Theorem B give a* $\mathbb{Z}[t]$*–basis for* $\mathrm{PH}_*(\mathrm{no}_*(n, \mathcal{W}))$.

Proof Recall that a basic cycle consists of a product of single vertices corresponding to singletons and boundaries of permutohedral cells corresponding to leader–follower pairs.

The key fact is that, as we increase w, a particular cell stays critical as long as, for every leader–follower pair, the weights add up to at least $w + 1$. But, once they add up to only w for some leader–follower pair, that means that the corresponding permutohedron boundary is filled in, and so the cycle becomes a boundary. Therefore, every cell of $P(n)$ which is critical in $P(n, \mathcal{W}, w)$ for some w corresponds to a direct summand of the persistence module. □

Proof of Theorem 7.1 Theorem 7.1 follows easily from Theorem 7.2. This is because the splitting

$$H_*(\mathrm{cell}(n, w)) = \bigoplus_{\sigma \in S_n} H_{*-\#\sigma}\big(P(n - \#\sigma, \mathcal{W}(\sigma), w)\big)$$

of Theorem 4.1 is natural with respect to increasing w (even on the chain level, as one readily sees). Therefore,

$$\mathrm{PH}_*(\mathrm{cell}(n, *)) = \bigoplus_{\sigma \in S_n} \mathrm{PH}_{*-\#\sigma}\big(P(n - \#\sigma, \mathcal{W}(\sigma), *)\big).$$

In particular, an element leaves the basis exactly when one of its filters has the wheel sizes adding up to at most w, and the correspondence with the no-$(w+1)$-equal homology proves that the element is null-homologous in this case.

For cohomology, the restriction map $H^*(\mathrm{config}(n, w)) \to H^*(\mathrm{config}(n, w-1))$ corresponds to intersecting each basis element $V(g)$ of $H_{2n-*}(M(n, w), \partial M(n, w))$ with the smaller space $M(n, w-1)$. When w gets too small for $V(g)$ to be in the basis, it is because some block of g has more than w elements; thus, the intersection of $V(g)$ with this $M(n, w)$ is empty, and the restricted cohomology class is zero. Thus, it is also true for cohomology that, when a given element stops being in the basis, it becomes zero. □

Remark 7.3 We can also explore how the persistence module structure on homology and cohomology interacts with other structures we have discussed:

(1) Cohomological persistence does not play nicely with the cup product structure: frequently a class is born indecomposable at time w and becomes decomposable at time $w - 1$. For example, in the notation of the previous section, we have the relation

$$\nu(5\,4\,|\,6\,2\,3\,|\,1) \cup \nu(4\,1\,|\,6\,|\,5\,|\,3\,|\,2) = \nu(5\,4\,1\,|\,6\,2\,3)$$

in $\mathrm{config}(6, 3)$ and $\mathrm{config}(6, 4)$, but $\nu(5\,4\,1\,|\,6\,2\,3)$ is indecomposable in $\mathrm{config}(6, 5)$.

(2) The concatenation product is perfectly well defined on persistence modules. So we can think of $\mathrm{PH}_*(\mathrm{config}(-, *))$ as a graded twisted algebra in $\mathbb{Z}[t]$–modules! The $\mathbb{Z}[t]$–module structure of this algebra is relatively easy to understand, as we have seen above, but it does not interact in a nice way with the symmetric group action. Moreover, unlike the algebras $H_*(\mathrm{config}(-, w))$ for fixed w, it is not finitely generated as an algebra. For these reasons, we do not study this structure further.

7.1 Asymptotics of the persistence module

A main goal of [Alpert et al. 2021] was to understand the growth of Betti numbers of $\mathrm{config}(n, w)$ as n increases. Now that we have described the persistence module of $\mathrm{config}(n, *)$, we can refine this: as the number of disks increases, we keep track of the number of bars of different lengths. It turns out that the number of short bars grows faster and eventually dominates the number of longer bars. We make this precise in the following theorem:

Theorem 7.4 *Each $\mathbb{Z}[t]$–basis element of $\mathrm{PH}_*(\mathrm{config}(n, *))$ born at $w = w_0$ either persists for all $w > w_0$ or dies by $w = 2w_0$. For each j and w_0, either the maps*

$$H_j(\mathrm{config}(n, w_0)) \to H_j(\mathrm{config}(n, w))$$

are isomorphisms for all $w > w_0$ and all n, or the fraction of basis elements of $H_j(\mathrm{config}(n, w_0))$ that persist to $H_j(\mathrm{config}(n, w_0 + 1))$ approaches 0 as n approaches ∞.

We note that, because each homology basis element is matched to a cohomology basis element in the same degree with the same birth and death, the theorem for homology immediately implies a corresponding statement for cohomology, which we do not include here.

For the second statement of Theorem 7.4, it helps to have an asymptotic estimate of the dimension of $H_j(\mathrm{config}(n, w))$ as n approaches ∞. Examining the basis for homology and estimating the number of elements recovers the following theorem:

Theorem 7.5 [Alpert et al. 2021] *If $w \geq 2$ and $0 \leq j \leq w - 2$, then the inclusion of $\mathrm{config}(n, w)$ into the configuration space of points in the plane induces an isomorphism on H_j. If $w \geq 2$ and $j \geq w - 1$, then there are positive constants c_1 and c_2, depending on w and j, such that the following is true. Write $j = q(w - 1) + r$ with $q \geq 1$ and $0 \leq r < w - 1$. Then*

$$c_1 \cdot n^{qw+2r}(q + 1)^n \leq \dim H_j(\mathrm{config}(n, w)) \leq c_2 \cdot n^{qw+2r}(q + 1)^n.$$

Proof of Theorem 7.4 The basis elements that persist indefinitely are those with no filters. Each filter can be written as a leader–follower pair, and any leader–follower pair that appears at time w_0 has total weight at most $2w_0$, because the leader block has at most the weight of the follower block. Thus, it only remains a filter until at most $w = 2w_0$. This proves the first statement of the theorem.

For the second statement, by Theorem 7.5 it suffices to show that, for $j \geq w - 1$, the number of basis elements of $H_j(\mathrm{config}(n, w))$ that persist to $H_j(\mathrm{config}(n, w + 1))$ grows polynomially in n. We observe that, if a filter for w contains any wheel of size 1, then its total size is $w + 1$, so it does not remain a filter for $w + 1$. Thus, in any basis element of $H_j(\mathrm{config}(n, w))$ that persists, every filter must contain only wheels of size at least 2. One rule for being a basis element is "Every wheel immediately to the left of a filter is greater than the least wheel in the filter", so this then implies that, in any basis element of $H_j(\mathrm{config}(n, w))$ that persists, all of the wheels of size 1 appear to the right of all other wheels and filters.

To estimate the number of such basis elements, we simply count the cell symbols that end with (at least) $n - 2j$ singleton blocks in descending order. There are $\binom{n}{2j} \cdot (2j)! \cdot 2^{2j-1}$ such symbols, and that function grows polynomially in n. Because the total number of basis elements of $H_j(\mathrm{config}(n, w))$ grows exponentially in n, asymptotically almost all of the basis elements do not persist to $w + 1$. $\square$

8 Relations in the twisted algebra and FI_d–modules

The twisted algebra structure of $H_*(\mathrm{cell}(-, w))$ is unusual because many pairs of elements do not commute. In particular, there are some elements that do not commute with the 0–cycles coming from a single disk; we think of these as *barriers* that prevent singleton disks from passing back and forth. This noncommuting with singleton 0–cycles is the main reason for Theorem 7.5: a given homology element in degree j can be written as the concatenation product of up to $\lfloor j/(w-1) \rfloor$ barriers, giving up to $1 + \lfloor j/(w-1) \rfloor$ nonequivalent ways to insert a new disk as a singleton.

One algebraic object that exhibits this kind of exponential growth is an FI_d–module. The best example for understanding the idea of an FI_d–module is the jth homology of the configuration space of n disks on the disjoint union of d planes. Each additional disk can be added to any of the d planes. Sam and Snowden [2017] define FI_d–modules for the first time. Ramos [2017] shows that finitely generated FI_d–modules satisfy a notion of generalized representation stability, and [2019] that the homology groups of a certain kind of graph configuration space are finitely generated FI_d–modules.

We show in this section that the homology of unweighted no-$(w+1)$-equal spaces and of $\mathrm{config}(n, 2)$ are both FI_d–modules. On the other hand, we give an example to show that the homology of $\mathrm{config}(n, w)$ for $w > 2$ is probably not well described via FI_d–modules, since there seems to be no consistent way of decomposing homology classes as products of barriers.

Formally, the category FI_d has one object $[n] = \{1, \ldots, n\}$ for each natural number n. The morphisms are pairs (φ, c), where φ is an injection, say from $[n]$ to $[m]$, and c is a d–coloring on the complement of
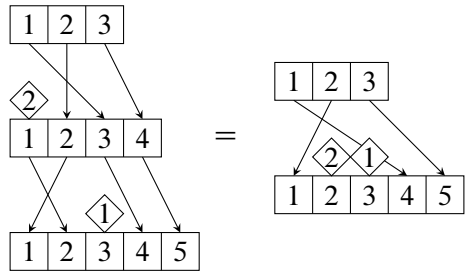
Figure 11: To compose two morphisms in FI_d, we have $(\varphi', c') \circ (\varphi, c) = (\varphi' \circ \varphi, c'')$, where $c''(i)$ is equal to $c'(i)$ if i is not in the image of φ' (for instance, $i = 3$ has color 1 in the example shown) and is equal to $c(\varphi'^{-1}(i))$ if i is in the image of φ' (for instance, $i = 2$ has color 2 in the composition because $c(1) = 2$ and $\varphi'(1) = 2$).

the image of φ; that is, c is a map from $[m] \setminus \varphi([n])$ to a set of size d, which in this paper we choose to be $\{0, 1, \ldots, d - 1\}$. The morphisms compose as illustrated in Figure 11: for each element colored by the first morphism, in the composition, the image of that element under the second morphism is the one that gets that color. (In the picture, the color of a given element is shown in a diamond just above the element.) More formally, if $(\varphi, c) : [n_1] \to [n_2]$ and $(\varphi', c') : [n_2] \to [n_3]$ are two morphisms, then we have

$$(\varphi', c') \circ (\varphi, c) = (\varphi' \circ \varphi, c''),$$

where $c''(i)$ is equal to $c'(i)$ if $i \notin \varphi'([n_2])$, and is equal to $c(\varphi'^{-1}(i))$ if $i \in \varphi'([n_2])$.

An FI_d–*module* M over a commutative ring k is defined to be a functor from FI_d to k–modules; that is, we have a k–module M_n for each n, and for each $(\varphi, c) : [n] \to [m]$ we have a corresponding k–module map $(\varphi, c)_* : M_n \to M_m$. Here we use $k = \mathbb{Z}$. An FI_d–module is *finitely generated* if there exists a finite set of elements $x_1, \ldots, x_r \in \bigsqcup_{n=1}^{\infty} M_n$ such that the only FI_d–submodule of M containing $x_1, \ldots, x_r$ is M itself. Figure 12 sketches the FI_{j+1}–module structure for $H_j(\mathrm{config}(n, 2))$; the colors of the disks, shown in the picture as the numbers in the diamonds, indicate where to insert the disks between barriers.
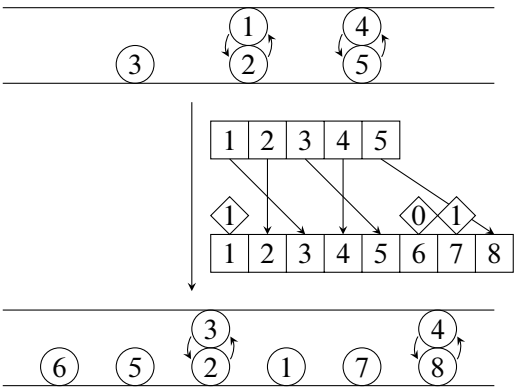

Figure 12: When applying this FI_3–morphism to a class in $H_2(\mathrm{config}(5, 2))$, we insert the disks with color k labels immediately after the kth circling pair.

8.1 No-$(w+1)$-equal spaces

Theorem 8.1 *The graded twisted algebra $H_*(\mathrm{no}_{w+1}(-))$ has a presentation with two generators and two relations. The generators are the representations spanned by a singleton 0–cycle and the $(w-1)$–cycle $\partial(1\,2\cdots w+1)$; these are trivial representations of S_1 and S_{w+1}, respectively. The relations are that two singletons commute and, for each $(w+2)$–block, taking the signed sum of the boundaries of those facets that have a $(w+1)$–block in them gives zero.*

Proof The fact that our proposed generators do generate comes automatically from the description of the basis. To show that the relations are true, we know that $\partial^2 = 0$, in particular when applied to a $(w+2)$–block, so the signed sum of the boundaries of *all* facets of the $(w+2)$–block gives zero. The facets that have no $(w+1)$–block in them are cells in $P(n, w)$, so their boundaries are null-homologous; thus, the sum of the boundaries of the remaining facets gives zero in homology.

To show that the specified relations are sufficient, we take every product of generators that is not in the basis and use the specified relations to write it in terms of the basis. First we may assume that consecutive singletons are always in descending order, using the relation that commutes singletons. Then, if a generator is not in the basis, some boundary of a $(w+1)$–block must be immediately preceded by a singleton that is less than every element of the $(w+1)$–block. If we combine those elements to form a $(w+2)$–block, this substring of our generator is the boundary of one facet of the $(w+2)$–block, and the boundaries of all the other facets are in the basis. Thus, applying the relation replaces a nonbasis substring of our generator by a sum of basis substrings. Applying the relations repeatedly from left to right rewrites our original nonbasis generator in terms of the basis. □

Theorem 8.2 *For each j a multiple of $w - 1$, the homology groups $H_j(\mathrm{no}_{w+1}(-))$ form a finitely generated FI_d–module for $d = 1 + j/(w - 1)$.*

Proof The FI_d–module structure is as follows. Suppose we have an FI_d–morphism from $[m]$ to $[n]$ with $n \geq m$. For any element of $H_j(\mathrm{no}_{w+1}(m))$, we write it in terms of the generators from Theorem 8.1. We apply the relabeling injection and, for each additional number with color $i \in \{0, 1, \ldots, j/(w - 1)\}$, we insert an element of that color into each summand between the ith and the $(i+1)^{\mathrm{st}}$ factors of the generator that look like the boundary of a $(w+1)$–block (rather than a singleton). The result is an element of $H_j(\mathrm{no}_{w+1}(n))$.

To show that this map on homology is well defined, we need to show that, if we write an element of $H_j(\mathrm{no}_{w+1}(m))$ in terms of the generators in two different ways, the resulting elements of $H_j(\mathrm{no}_{w+1}(n))$ are the same. To see this, suppose that we apply a relation to an original generator. The same relation can just as easily be applied *after* the relabeling and insertion, just by moving the new singletons past old singletons so that they are out of the way. Thus, applying the relation does not change which homology class we get.

The fact that the maps respect composition of FI_d–morphisms is automatic once we use the fact that consecutive singletons commute. Thus, we have an FI_d–module. It is finitely generated by the basis of $H_j\big(\mathrm{no}_{w+1}(j(w+1))\big)$. $\qquad\square$

A proof along similar lines shows that the homology of weighted no-$(w+1)$-equal spaces can be written as a direct sum, with each summand equal to the span of the generators with a particular number of filters. This allows us to insert weight-1 points between the filters in a well-defined way. However, the result is not formally an FI_d–module (or a direct sum of them) in a reasonable way, because relabeling cannot permute points of different weights without leaving the no-$(w+1)$-equal space. Because the statement of the theorem would be cumbersome, we do not include the details.

8.2 Disks in a strip of width $w = 2$

Essentially the same proofs as for the no-$(w+1)$-equal spaces give generators and relations for the configuration spaces for $w = 2$, which then give its homology an FI_d–module structure.

Theorem 8.3 *The graded twisted algebra $H_*(\mathrm{config}(-, 2))$ has a presentation with three generators and three relations. The generators are*:

(1) $H_0(\mathrm{config}(1, 2)) \cong \mathbb{Z}$, *ie a singleton 0–cycle on which S_1 acts trivially.*

(2) $H_1(\mathrm{config}(2, 2)) \cong \mathbb{Z}$, *ie a 2–wheel on which S_2 acts trivially. Write this as $w(1, 2) = 2\,1 + 1\,2$.*

(3) *The two-dimensional representation of S_3 spanned by the cycles*

$$z(1, 2, 3) = \overline{\raisebox{0pt}{$\begin{smallmatrix}2\\1\ 3\end{smallmatrix}$}} \quad and \quad z(1, 3, 2) = \overline{\raisebox{0pt}{$\begin{smallmatrix}1\ 3\\2\end{smallmatrix}$}}$$

in $H_1(\mathrm{config}(3, 2))$, where transpositions in S_3 act by switching the two basis vectors. This representation is irreducible over $\mathbb{Z}$, but over $\mathbb{Q}$ it splits into the direct sum of a trivial representation and a sign representation.

The relations are:

(1) *The singletons commute*: $a \,|\, b = b \,|\, a$.

(2) *The relation induced by boundaries of 4–cells in* $\mathrm{cell}(4)$.

(3) *The relation, in $H_1(\mathrm{cell}(3, 2))$,*

$$z(1, 2, 3) + z(1, 3, 2) = 1\,|\,w(2, 3) + w(2, 3)\,|\,1 + 2\,|\,w(3, 1) + w(3, 1)\,|\,2 + 3\,|\,w(1, 2) + w(1, 2)\,|\,3.$$

Rationally, one can use the sign representation in $H_1(\mathrm{config}(3, 2))$ as a generator, removing the need for the third relation.

Proof We first establish a basis for $H_*(\mathrm{config}(n, 2))$ consisting of concatenation products of (1), (2) and (3). We use the same critical cells as in Remark 4.5, but interpret them differently: leader–follower pairs of the form $a\,|\,b\,c$ where $a < b < c$ become $z(a, b, c)$ factors, while the rest of the blocks individually yield singletons and 2–wheels.

To show that the generators and relations are valid, it suffices to show:

(i) Any concatenation product of generators can be written as a sum of basis elements by rewriting via the relations.

(ii) Every element of the basis from Theorem B can be written as a sum of concatenation products of the new generators.

Together, the two steps imply that the new basis spans $H_*(\mathrm{config}(n, w))$; since it has the same cardinality as the old basis, this shows that it is indeed a basis. The first step then implies that the provided relations are sufficient.

To see (ii), we note that

(8.4) $$\partial(a \, b \, c) + z(a, b, c) = w(b, c) \, | \, a + b \, | \, w(c, a) + w(a, b) \, | \, c.$$

In this fashion we obtain filters of singletons. We already know that every filter of a singleton and a 2–wheel is given by an expression of the form

$$a \, | \, w(b, c) - w(b, c) \, | \, a.$$

Every element of the old basis is a concatenation product of these two types of cycles as well as singletons and 2–wheels.

To see (i), note first that, in the description of the basis, the 2–wheels separate the strip into intervals that do not interact with each other or with the 2–wheels; that is, the requirements for being a basis element are the same as the requirements for each of these intervals individually to give a basis element. Then, to rewrite a product of generators in terms of basis elements, we only need to rewrite products of singletons and 3–disk generators.

Relation (3) allows us to eliminate 3–disk generators that are in the "wrong" order. Equation (8.4) lets us turn relation (2) into a way to eliminate subwords of the form $a \, | \, z(b, c, d)$ where $a < b < c < d$. Thus, we can rewrite every product of singletons and 3–disk generators in normal form using the same method as in Theorem 8.1. □

Theorem 8.5 *For each* j, *the homology groups* $H_j(\mathrm{config}(-, 2))$ *form a finitely generated* FI_{j+1}–*module.*

Proof Every generator of $H_j(\mathrm{config}(n, 2))$ has exactly j factors of types (2) and (3). We define the FI_{j+1}–module structure as follows: to insert a new disk of color $i \in \{0, 1, \ldots, j\}$ into a generating cycle, we insert it as a singleton 0–cycle between the ith and $(i+1)^{\mathrm{st}}$ factors of type (2) or (3) of the generator.

As in Theorem 8.2, the relations can be applied either before or after the insertion, so the FI_{j+1}–module structure is well defined. The basis of $H_j(\mathrm{config}(3j, 1))$ gives a finite generating set for the FI_{j+1}–module. □

8.3 Disks in a strip of width $w > 2$

We say that an element of $H_*(\text{config}(n, w))$ is a *barrier* if the two ways to concatenate it with a 1–disk 0–cycle represent different homology classes. The wheels of size w are barriers, as are any filters that contain a wheel of size 1. Because the homology $H_*(\text{config}(n, w))$ is generated by concatenations of wheels and filters, we can count the number of barriers in each generator.

The following proposition implies that counting barriers is not well defined on arbitrary homology classes in $H_*(\text{config}(n, w))$. As mentioned above, this suggests that the structure of $H_*(\text{config}(n, w))$ is not well described by FI_d–modules.

Proposition 8.6 *There is a nontrivial element of $H_4(\text{config}(8, 3))$ that is simultaneously a sum of nonbarrier generators, a sum of one-barrier generators, and a sum of two-barrier generators.*

As a warm-up before proving the proposition, we consider the case of $n = 4$, $w = 3$ and $\sigma = 2\ 1\ 4\ 3$. Then $P(n - \#\sigma, \mathcal{W}(\sigma), w)$ is a no-$(w+1)$-equal space with two points of weight 2, which we denote by [2 1] and [4 3]. Then, because i_σ is a chain map, we have

$$i_\sigma\big(\partial([2\ 1][4\ 3])\big) = \partial\big(i_\sigma([2\ 1][4\ 3])\big).$$

The left-hand side is the sum (with some signs) of the two ways to concatenate the wheels 2 1 and 4 3, whereas the right-hand side is the sum (with some signs) of the four cycles $\partial(2\ 1\ 4\ 3)$, $\partial(2\ 1\ 3\ 4)$, $\partial(1\ 2\ 4\ 3)$, and $\partial(1\ 2\ 3\ 4)$, each of which can be thought of as an image under the S_4–relabeling of $\partial(1\ 2\ 3\ 4)$, which, for width $w = 3$, is a filter with four wheels each containing one disk, and is a barrier.

Proof of Proposition 8.6 Let $\sigma = 2\ 1\ 4\ 3\ 6\ 5\ 8\ 7$. Then $P(n - \#\sigma, \mathcal{W}(\sigma), w)$ is a no-$(w+1)$-equal space with four points of weight 2, which we denote by [2 1], [4 3], [6 5] and [8 7]. We can apply any element $\tau \in S_4$ to any chain in $C_*\big(P(n - \#\sigma, \mathcal{W}(\sigma), w)\big)$ by permuting these four labels of points.

To construct our cycle, we take

$$z = \sum_{\tau \in S_4} \text{sign}(\tau) \cdot i_\sigma\big(\tau([2\ 1]\,|\,[4\ 3]\,|\,[6\ 5]\,|\,[8\ 7])\big).$$

That is, we take the signed sum of all of the ways to concatenate the four wheels 2 1, 4 3, 6 5 and 8 7. Each of these summands is already a basis element, although they differ as to which pairs are considered filters of two wheels. Thus, their sum is nonzero in homology, and none of these wheels is a barrier, so it is a sum of nonbarrier generators.

To write our element z as a sum of one-barrier cycles, we pair up all elements τ in S_4 that differ by swapping the middle two wheels. One such pair gives

$$i_\sigma\big([2\ 1]\,|\,\partial([4\ 3][6\ 5])\,|\,[8\ 7]\big),$$

which concatenates the wheels 2 1 (on the left) and 8 7 (on the right) to the cycle

$$i_\sigma\big(\partial([4\,3][6\,5])\big) = \partial\big(i_\sigma([4\,3][6\,5])\big),$$

which is a sum (with some signs) of relabelings of the barrier filter $\partial(3\,4\,5\,6)$. In this way we can write our element z as a sum of generators, each one a concatenation of a 2–wheel, a barrier filter, and another 2–wheel.

To write our element z as a sum of two-barrier cycles, we group the elements τ in S_4 into quadruples: permutations get grouped together if they differ by swapping the first two and/or the last two wheels. One such quadruple gives

$$i_\sigma\big(\partial([2\,1][4\,3]) \mid \partial([6\,5][8\,7])\big) = \pm\partial\big(i_\sigma([2\,1][4\,3])\big) \mid \partial\big(i_\sigma([6\,5][8\,7])\big),$$

which, similar to the computation above, is a sum of generators, each one a concatenation of two barrier filters. □

9 Configuration spaces of unordered disks

The configuration space of n unordered disks of diameter 1 in a strip of width w is the quotient of config(n, w) by the action of S_n that permutes the disk labels, and it is homotopy equivalent to the quotient of cell(n, w) by the action of S_n. Because the action is cellular and free, this quotient is a cell complex, which we call ucell(n, w). In this section, we compute the homology of ucell(n, w) (and thus of the configuration space of unordered disks in a strip) with field coefficients, using the discrete Morse theory methods from Section 3. In the version of discrete Morse theory that applies here, we do not need to assume that the coefficients of the boundary map are ± 1, as is true for polyhedral cell complexes, but only that these coefficients are units. We use field coefficients throughout this section, so all nonzero coefficients are automatically units.

The concatenation product

$$H_j(\text{ucell}(n, w)) \otimes H_{j'}(\text{ucell}(n', w)) \to H_{j+j'}(\text{ucell}(n + n', w))$$

is well defined since there is no need to choose labels for the disks. Therefore, for any ring R, $H_*(\text{ucell}(*, w); R)$ forms a noncommutative bigraded R–algebra.

The cells of ucell(n, w) are labeled by symbols as in cell(n, w), but the numbers in the symbols become indistinguishable; the only remaining information is the sizes of blocks, which we also refer to as weights. We notate them all by ∘, and use exponents to denote the weights of the blocks. For instance, in cell(3) we have

$$\partial(1\,2\,3) = -1 \mid 2\,3 + 2 \mid 1\,3 - 3 \mid 1\,2 + 2\,3 \mid 1 - 1\,3 \mid 2 + 1\,2 \mid 3,$$

and the corresponding relation in ucell(3) is

$$\partial(\circ^3) = -\circ \mid \circ\circ + \circ \mid \circ\circ - \circ \mid \circ\circ + \circ\circ \mid \circ - \circ\circ \mid \circ + \circ\circ \mid \circ = -\circ \mid \circ\circ + \circ\circ \mid \circ.$$

The following lemma describes all the coefficients of these boundary maps:

Lemma 9.1 *In* ucell(n), *the coefficient of the face* $\circ^k \mid \circ^{n-k}$ *in the boundary of the cell* $\circ^n$ *can be described as follows*:

- *If k and $n-k$ are both odd, the coefficient is 0.*
- *If $n = 2n'$ is even and $k = 2k'$ is even, the coefficient is $\binom{n'}{k'}$.*
- *If $n = 2n' + 1$ is odd and $k = 2k'$ is even, the coefficient is $\binom{n'}{k'}$.*
- *If $n = 2n' + 1$ is odd and $k = 2k' + 1$ is odd, the coefficient is $-\binom{n'}{k'}$.*

Proof Consider the cell $1\, 2 \cdots n$ in cell(n). In its boundary, there are $\binom{n}{k}$ cells that project to $\circ^k \mid \circ^{n-k}$ in ucell(n), and our task is to add up all of their signs. The sign of each such face is $(-1)^k$ times the sign of the corresponding permutation.

We pair up the numbers 1 and 2, 3 and 4, and so on, pairing up $n-1$ and n if n is even, or leaving only n unpaired if n is odd. We can match and cancel faces of $1\, 2 \cdots n$ with opposite signs in the following way. Given a face, if 1 and 2 are in different blocks, then swapping them gives another face with opposite sign. Similarly, if 1 and 2 are in the same block, but 3 and 4 are in different blocks, then swapping 3 and 4 gives another face with opposite sign. In this way, for each face for which a pair of numbers is split up, we match and cancel it with another such face by finding the first pair of numbers that is split up, and swapping those numbers.

The remaining faces have 1 and 2 in the same block, 3 and 4 in the same block, and so on, and each one corresponds to an even permutation, so the total sign is $(-1)^k$. If k and $n-k$ are both odd, then there are no such faces. If $k = 2k'$ is even and there are n' pairs, then the faces all have positive sign, and there are $\binom{n'}{k'}$ of them. And, if $k = 2k' + 1$ is odd and there are n' pairs, then the faces all have negative sign, and there are $\binom{n'}{k'}$ of them. $\square$

9.1 Discrete Morse theory on ucell(n, w) with $\mathbb{Q}$ coefficients

When ordering the cells to produce a discrete gradient vector field, part of the ordering will be chosen later to be lexicographical. Thus, we need to compute the lexicographically least way to split each block.

Lemma 9.2 *The lexicographically least face of the cell $\circ^n$ in* ucell(n) *with nonzero coefficient is given as follows. If n is odd, the least face is $\circ^1 \mid \circ^{n-1}$. If n is even and greater than 2, the least face is $\circ^2 \mid \circ^{n-2}$. In the case $n = 2$, the cell $\circ^2$ is a cycle.*

Proof If n is odd, then the coefficient of the face $\circ^1 \mid \circ^{n-1}$ is $-\binom{(n-1)/2}{0} = -1$. If $n > 2$ is even, then the coefficient of the face $\circ^1 \mid \circ^{n-1}$ is zero, but the coefficient of the face $\circ^2 \mid \circ^{n-2}$ is $\binom{(n-2)/2}{1} = \frac{1}{2}(n-2) \neq 0$. $\square$

Definition Two consecutive blocks in a symbol in ucell(n, w) form a *leader–follower pair in characteristic* 0 if they have the form $\circ^1 \mid \circ^{2k'}$ or $\circ^2 \mid \circ^{2k'}$ for some k'.

Theorem 9.3 *There is a discrete gradient vector field on* $\mathrm{ucell}(n, w)$ *such that the critical cells are in bijection with a basis for* $H_*(\mathrm{ucell}(n, w); \mathbb{Q})$. *The critical cells are described as follows*:

- *For* $w = 2$, *every cell.*
- *For* $w = 3$, *the concatenation of zero or more blocks* $\circ^2$, *followed by zero or more singletons* $\circ^1$.
- *For* $w > 2$ *even, the concatenation of zero or more copies of* $\circ^2 \,|\, \circ^w$ *or strings ending in* $\circ^1 \,|\, \circ^w$, *followed by the concatenation of zero or more singletons* $\circ^1$. *For each string ending in* $\circ^1 \,|\, \circ^w$, *it consists of an optional* $\circ^2$, *followed by zero or more singletons* $\circ^1$, *followed by the pair* $\circ^1 \,|\, \circ^w$.
- *For* $w > 3$ *odd, the concatenation of zero or more pairs* $\circ^2 \,|\, \circ^{w-1}$, *followed by an optional* $\circ^2$, *followed by zero or more singletons* $\circ^1$.

Proof The proof is similar to the proof of Theorem 3.4, which finds a basis for homology of weighted no-$(w+1)$-equal spaces. Informally, we construct a discrete vector field that pairs cells as follows: given a symbol, we read it from left to right, and find the first place where either there is a block of weight greater than 2, in which case we match down by breaking it into a leader–follower pair; or there is a leader–follower pair of total weight at most w, in which case we match up by merging it into one block.

The cells that remain unpaired are those for which every leader–follower pair has total weight greater than w and all other blocks are either $\circ^1$ or $\circ^2$. We observe that, if w is even, the only possibilities for leader–follower pairs in critical cells are $\circ^1 \,|\, \circ^w$ and $\circ^2 \,|\, \circ^w$, because the total weight must exceed w while the weight of the follower block must be even and at most w. For the same reason, if w is odd, the only possible leader–follower pair in a critical cell is $\circ^2 \,|\, \circ^{w-1}$. We also observe that there are no instances of either $\circ^1 \,|\, \circ^2$ for $w \geq 3$, or $\circ^2 \,|\, \circ^2$ for $w \geq 4$, because these would be leader–follower pairs. Combining these observations, we deduce that the critical cells must have the form given in the theorem statement.

More formally, to check that this discrete vector field is gradient, we exhibit an ordering that produces it. We order the cells of $\mathrm{ucell}(n, w)$ as follows: if f and g are two cells, we find the first block where they differ, and order according to this first differing block in f and g:

- a follower block is less than a nonfollower;
- two follower blocks are ordered in increasing order of weight; and
- two nonfollower blocks are ordered in decreasing order of weight.

Then we pair up two cells f and g, with f a face of g, if f is the greatest face of g and g is the least coface of f. One can check that this pairing agrees with the pairing described informally above, and thus that the critical cells match the desired description.

To check that the resulting critical cells correspond to a basis, by Lemma 3.1 it suffices to find a cycle $z(e)$ for each critical cell e such that e is the greatest cell in $z(e)$. Because we are using coefficients in $\mathbb{Q}$, we

do not have to worry about whether the coefficient is a unit, as long as it is nonzero. To construct $z(e)$, we take the concatenation product of the following cycles: for each block that is not in a leader–follower pair, it is already a cycle, and, for each leader–follower pair, as our cycle we take the boundary of the block resulting from merging the pair. Note that every cell of this boundary that has a nonzero coefficient has block weight at most w, because the original leader–follower pair is one of the two faces with the most unbalanced block weights in this boundary. Because our leader–follower pair is the lexicographically least face of the merged block, and our ordering is the reverse of lexicographical for nonfollowers, we see that e is the greatest cell in $z(e)$. This implies that the cells $z(e)$ form a basis for $H_*(\mathrm{ucell}(n, w); \mathbb{Q})$. $\quad\square$

Corollary 9.4 *The homology $H_*(\mathrm{ucell}(*, w); \mathbb{Q})$ forms a bigraded algebra over $\mathbb{Q}$ under concatenation product. It has the following generators*:

(1) *the singleton block $\circ^1$;*

(2) *the block $\circ^2$;*

(3) *for $w > 3$, the cycle $\partial(\circ^{w+1})$; and*

(4) *for $w > 2$ even, the cycle $\partial(\circ^{w+2})$.*

It has the following relations:

(1) *the singleton block $\circ^1$ commutes with $\circ^2$ for all $w \geq 3$;*

(2) *the singleton block $\circ^1$ commutes with $\partial(\circ^{w+1})$ for all odd $w > 3$; and*

(3) *the symbol $\circ^2 \mid \circ^2$ is null-homologous for all $w \geq 4$.*

Proof The description of the basis shows that the specified generators do generate.

Relation (1) is true because the boundary of the cell $\circ^3$ is $-\circ^1 \mid \circ^2 + \circ^2 \mid \circ^1$. Relation (2) comes from the relation $\partial^2(\circ^{w+2}) = 0$ on $\mathrm{ucell}(n)$; when w is odd, expanding $\partial(\circ^{w+2})$ gives $-\circ^1 \mid \circ^{w+1} + \circ^{w+1} \mid \circ^1$ plus a sum of cells in $\mathrm{ucell}(n, w)$, so applying ∂ again gives our desired homology relation. Relation (3) is true because $\circ^2 \mid \circ^2$ is the only face of $\circ^4$ with nonzero coefficient.

The relations are enough to transform an arbitrary product of generators into one of our basis cycles. $\quad\square$

Corollary 9.5 *For fixed j and w, the Betti numbers $\beta_j(\mathrm{ucell}(n, w); \mathbb{Q})$ as a function of n grow with an upper bound of $O(n^q)$, where $q = \lfloor j/(w-1) \rfloor$. If w is odd, the Betti numbers either are 0 for all n or are 1 for all sufficiently large n. If w is even, the Betti numbers either are 0 for all n or grow as $\Theta(n^q)$, and the latter case holds for all $j \geq (w-1)(w-3)$.*

Proof Deleting nonleader singleton blocks $\circ^1$ from a critical cell gives a critical cell with smaller n, and, for each j and w, there are finitely many ways to form one of these "skyline" critical cells that have no nonleader singletons. For each skyline critical cell, the only places to insert singletons are at the end (that

is, on the right side) and immediately before the leader–follower pair $\circ^1 \mid \circ^w$, which exists only if w is even. Thus, if w is odd, then, for all sufficiently large n, the Betti number $\beta_j(\text{ucell}(n, w); \mathbb{Q})$ is constant, equal to the number of skyline critical cells for j and w, which is 1 if j is congruent to 0 or 1 mod $w - 1$, and 0 otherwise.

If w is even, for each j we claim that either there is a skyline with $\lfloor j/(w-1) \rfloor$ instances of $\circ^1 \mid \circ^w$, or there is no skyline at all. We write j as $q(w-1) + r$, with $0 \leq r \leq w - 2$. To construct the critical cell, if $r \leq q$, we concatenate r instances of $\circ^2 \mid \circ^1 \mid \circ^w$ and then $q - r$ instances of $\circ^1 \mid \circ^w$. If $r = q + 1$, we concatenate q instances of $\circ^2 \mid \circ^1 \mid \circ^w$, followed by one instance of $\circ^2$. If $r > q + 1$, there is no way to build a critical cell of dimension j, because all blocks that contribute to the dimension are either $\circ^w$, contributing $w - 1$, or $\circ^2$, contributing 1, and there can be at most $q + 1$ instances of $\circ^2$.

Given a skyline critical cell with n' disks and k instances of $\circ^1 \mid \circ^w$, the number of critical cells with n disks arising from this skyline is $\binom{n-n'+k}{k}$, corresponding to the number of ways to arrange $n - n'$ additional singletons and k dividers. This is a polynomial in n of degree k. If w is even, for each j either there is no skyline, or there is a skyline with $q = \lfloor j/(w-1) \rfloor$ instances of $\circ^1 \mid \circ^w$, which is the largest possible k. Thus, the Betti numbers grow like either 0 or $\Theta(n^q)$.

In the case where w is even and $j \geq (w-1)(w-3)$, the quotient q is at least $w - 3$ and the remainder r is at most $w - 2$, so the case $r > q + 1$ is impossible, and the Betti numbers grow like $\Theta(n^q)$. $\square$

9.2 Discrete Morse theory on ucell(n, w) with $\mathbb{F}_p$ coefficients

Using coefficients mod p, our strategy for computing the homology is the same as with $\mathbb{Q}$ coefficients, but the answer becomes more complicated because we need to account for divisibility of binomial coefficients.

Lemma 9.6 *For any prime p, the lexicographically least face of the cell $\circ^n$ in ucell(n) with coefficient not divisible by p is given as follows. If n is odd, the least face is $\circ^1 \mid \circ^{n-1}$. If n is even, then we write n as $2p^k \cdot a$, where a is not divisible by p, and the least face is $\circ^{2p^k} \mid \circ^{2p^k(a-1)}$.*

Proof If n is odd, then the coefficient of the face $\circ^1 \mid \circ^{n-1}$ is $-\binom{(n-1)/2}{0} = -1$, which is a unit in any $\mathbb{F}_p$.

If $n = 2p^k \cdot a$ is even, then the faces of $\circ^n$ have coefficients $\binom{p^k a}{k'}$ for various k'. These are the coefficients of $(x + y)^{p^k a} \equiv (x^{p^k} + y^{p^k})^a$ mod p, using the Frobenius homomorphism. These coefficients are 0 unless p^k divides k', and the coefficient $\binom{p^k a}{p^k}$ is congruent mod p to $\binom{a}{1} = a$, which is a unit in $\mathbb{F}_p$. $\square$

Definition Two consecutive blocks in a symbol in ucell(n, w) form a *leader–follower pair in characteristic p* if they have the form $\circ^1 \mid \circ^{2k'}$ for some k', or the form $\circ^{2p^k} \mid \circ^{2p^k(a-1)}$ for some $k \geq 0$ and a not divisible by p. We assign the pairs disjointly from left to right, so that, once a block is a follower in a pair with the previous block, it cannot also be a leader in a pair with the next block.

Theorem 9.7 *For each prime p, there is a discrete gradient vector field on $\mathrm{ucell}(n, w)$ with $\mathbb{F}_p$ coefficients such that the critical cells are in bijection with a basis for $H_*(\mathrm{ucell}(n, w); \mathbb{F}_p)$. The critical cells are those with the properties that every leader–follower pair has total weight greater than w, and every block that is not a follower is either $\circ^1$ or has the form $\circ^{2p^k}$ for some $k \geq 0$. These properties imply that consecutive blocks that are not followers appear in decreasing order of weight, weakly decreasing if $p = 2$ and strictly decreasing if $p \neq 2$, except for singleton blocks $\circ^1$ which may occur consecutively for any p.*

Proof The proof is exactly analogous to that of Theorem 9.3, which addresses the case of $\mathbb{Q}$ coefficients. As in that proof, our discrete vector field is informally described by reading each symbol from left to right, breaking down any block of weight other than 1 or $2p^k$ into a leader–follower pair, and combining any leader–follower pair of total weight at most w.

We can describe the resulting critical cells more concretely as follows. To find all possibilities for leader–follower pairs in critical cells, for each $k \geq 0$ such that $2p^k \leq w$, we find the least multiple of $2p^k$ greater than w. If this multiple is not divisible by $2p^{k+1}$, it has the form $2p^k a$ for a not divisible by p, and the leader–follower pair $\circ^{2p^k} \mid \circ^{2p^k(a-1)}$ may appear in a critical cell. We know that $2p^k(a - 1)$ is at most w, otherwise it would contradict the selection of $2p^k a$ as the least multiple of $2p^k$ greater than w. If the least multiple of $2p^k$ greater than w is divisible by $2p^{k+1}$, there is no leader–follower pair beginning with $2p^k$ that may appear in a critical cell. In addition, if w is even, the leader–follower pair $\circ^1 \mid \circ^w$ may appear in a critical cell.

For each critical cell, we can imagine dividing the symbol into strings, where the followers are the dividers. Each string consists of blocks $\circ^1$ and/or $\circ^{2p^k}$ for various $k \geq 0$, with constraints on the multiplicities and order because they may not form leader–follower pairs. The pair of blocks $\circ^{2p^k} \mid \circ^{2p^k}$ forms a leader–follower pair if $p \neq 2$, so it may not appear in one of these strings; if $p = 2$, it does not form a leader–follower pair, so it may appear. The pairs $\circ^1 \mid \circ^{2p^k}$ for $k \geq 0$ and $\circ^{2p^k} \mid \circ^{2p^l}$ for $k < l$ do form leader–follower pairs regardless of p, so they may not appear in one of these strings. Thus, if $p = 2$, each string is an arbitrary sequence of blocks $\circ^1$ and $\circ^{2p^k}$, any number of each, in weakly decreasing order. If $p \neq 2$, each string is an arbitrary sequence of any number of blocks $\circ^1$ and at most one of each block $\circ^{2p^k}$, in decreasing order. Note that, given a leader–follower pair, in the string preceding it, the blocks cannot have smaller weight than the leader, because of the condition that the entire string including the leader should be in decreasing order.

To complete the proof formally, we use the same ordering on cells of $\mathrm{ucell}(n, w)$ as in the proof of Theorem 9.3, except with the characteristic p definition of leader–follower pairs. The resulting discrete gradient vector field agrees with the informal description above. We define cycles $z(e)$ as in the proof of Theorem 9.3: for each leader–follower pair, we take the boundary of the block in $\mathrm{ucell}(n)$ resulting from merging the pair, while each block not in a leader–follower pair is already a cycle, and we take the concatenation product of all these cycles to get $z(e)$. Applying Lemma 3.1, we conclude that the cells $z(e)$ form a basis for $H_*(\mathrm{ucell}(n, w); \mathbb{F}_p)$. □

Corollary 9.8 *For each prime p, the homology $H_*(\mathrm{ucell}(*,w);\mathbb{F}_p)$ forms a bigraded algebra over $\mathbb{F}_p$ under concatenation product. It has the following generators:*

(1) *the singleton block $\circ^1$;*

(2) *the block $\circ^{2p^k}$ for each $k \geq 0$ with $2p^k \leq w$;*

(3) *if w is even but not equal to $2p^k$ for any $k \geq 0$, the cycle $\partial(\circ^{w+1})$; and*

(4) *the cycle $\partial(\circ^{n'})$, where n' is the least multiple of $2p^k$ greater than w, for each $k \geq 0$ such that $2p^k \leq \frac{1}{2}w$; in this case, we let $k(n')$ denote the power of p in the prime factorization of n'.*

It has the following relations:

(1) *the singleton block $\circ^1$ commutes with $\circ^{2p^k}$ whenever $1 + 2p^k \leq w$;*

(2) *the singleton block $\circ^1$ commutes with $\partial(\circ^{n'})$ whenever $1 + n' - 2p^{k(n')} \leq w$;*

(3) *if $p \neq 2$, then $\circ^{2p^k} \mid \circ^{2p^k}$ is null-homologous whenever $4p^k \leq w$;*

(4) *we have $\circ^{2p^l} \mid \circ^{2p^k} = -(\circ^{2p^k} \mid \circ^{2p^l})$ whenever $2p^l + 2p^k \leq w$ and $k \neq l$; and*

(5) *the block $\circ^{2p^l}$ commutes with $\partial(\circ^{n'})$ whenever $2p^l + n' - 2p^{k(n')} \leq w$.*

Proof The description of the basis shows that the specified generators do generate.

Relation (1) is true because the boundary of the cell $\circ^{1+2p^k}$ mod p is $-\circ^1 \mid \circ^{2p^k} + \circ^{2p^k} \mid \circ^1$. All other faces have coefficients of 0, because $\binom{p^k}{k'} \equiv 0 \bmod p$ unless k' is 0 or p^k. Relation (2) comes from the relation $\partial^2(\circ^{1+n'}) = 0$ on $\mathrm{ucell}(n)$; expanding $\partial(\circ^{1+n'})$ gives $-\circ^1 \mid \circ^{n'} + \circ^{n'} \mid \circ^1$ plus a sum of cells in $\mathrm{ucell}(n,w)$, so applying ∂ again gives our desired homology relation. Relation (3) is true because $\circ^{2p^k} \mid \circ^{2p^k}$ is the only face of $\circ^{4p^k}$ with nonzero coefficient mod p when $p \neq 2$. (If $p = 2$, then $\circ^{4p^k}$ is a cycle.) To see that relation (4) is true, if $l < k$, the faces of $\circ^{2p^l+2p^k}$ have coefficients $\binom{p^l(1+p^{k-l})}{p^l k'} \equiv \binom{1+p^{k-l}}{k'} \bmod p$ for various k'. The coefficients for $k' = 1$ and $k' = p^{k-l}$ are both 1 mod p and are the coefficients of $\circ^{2p^l} \mid \circ^{2p^k}$ and $\circ^{2p^k} \mid \circ^{2p^l}$. The other coefficients are all 0 mod p, because Pascal's triangle identity gives $\binom{1+p^{k-l}}{k'} = \binom{p^{k-l}}{k'-1} + \binom{p^{k-l}}{k'}$, which is 0 mod p unless k' is 0, 1, $p^{k-l} - 1$ or p^{k-l}. Relation (5) comes from the relation $\partial^2(\circ^{2p^l+n'}) = 0$; using similar reasoning to relation (4), we find that the only faces of $\partial(\circ^{2p^l+n'})$ with nonzero coefficient have the form $\circ^{k'} \mid \circ^{2p^l+n'-k'}$, where k' is congruent to either 0 or $2p^l \bmod 2p^k$. Thus, $\partial(\circ^{2p^l+n'})$ is $\circ^{2p^l} \mid \circ^{n'} + \circ^{n'} \mid \circ^{2p^l}$ plus a sum of cells in $\mathrm{ucell}(n,w)$, and then we may apply ∂ again, keeping in mind that the sign convention for the Leibniz rule gives a negative sign to the term $\circ^{2p^l} \mid \partial(\circ^{n'})$.

The relations are enough to transform an arbitrary product of generators so that consecutive nonfollower blocks are in decreasing order, strictly decreasing if $p \neq 2$. The resulting cycle is in our basis. □

Corollary 9.9 *For fixed j, w and prime p, the Betti numbers $\beta_j(\mathrm{ucell}(n,w);\mathbb{F}_p)$ as a function of n grow with an upper bound of $O(n^q)$, where $q = \lfloor j/(w-1) \rfloor$. If w is odd, the Betti numbers become eventually constant in n; if w is even and $j \geq (w-1)(w-3)$, the Betti numbers grow as $\Theta(n^q)$.*

Proof As in Corollary 9.5, we can delete nonleader singleton blocks $\circ^1$ from each critical cell to form one of finitely many "skylines". To recover all critical cells with a given skyline, we insert singletons $\circ^1$ either on the far right, or immediately preceding a leader–follower pair $\circ^1 | \circ^w$, which can only exist if w is even. In this case, the maximum possible number of such pairs is $q = \lfloor j/(w-1) \rfloor$, so, summing over all skylines, we find that the total Betti number is eventually equal to the number of skylines if w is odd, and is bounded above by $O(n^q)$ if w is even.

In the case where $j \geq (w-1)(w-3)$, we can construct a skyline critical cell with q instances of $\circ^1 | \circ^w$ in exactly the same way as in Corollary 9.5. The existence of such a skyline implies that $\beta_j(\mathrm{ucell}(n, w); \mathbb{F}_p)$ grows with a lower bound of $\Omega(n^q)$, matching the upper bound. $\qquad\square$

10 Open questions and further directions

(1) One generalization of the disks in a strip configuration spaces is the following. Let $E \to B$ be a locally trivial bundle, and consider the configuration space of n distinct points in E such that each fiber of the bundle may contain at most w of them. What do our methods say about the homology of this configuration space? We predict that, if some neighborhoods in B are one-dimensional, then the homology exhibits the same noncommutativity as that of disks in a strip, but that, if B is everywhere at least two-dimensional, then the homology of the configuration space is a finitely generated FI–module.

(2) The representation stability properties of the configuration space of n points on a given manifold come in some sense from the special case where the manifold is Euclidean space; the special case can be considered a local model. In particular, when the manifold has an end, the homology of the manifold configuration space, considered for all n at once, is a module over the twisted commutative algebra given by the homology of the Euclidean configuration space. The algebra acts by inserting cycles near infinity in the end of the manifold. Does our disks-in-a-strip configuration space act as a local model for other configuration spaces, for which the homology exhibits similar finite generation properties due to its being a module? For instance, what can we say about the homology of the configuration space of disks in the product of an interval with a noncompact 1–complex, such as the union of three rays with a common starting point?

(3) Our proof that $H_*(\mathrm{config}(n, w))$ is finitely generated as a twisted noncommutative algebra relies on fully computing $H_*(\mathrm{config}(n, w))$ and exhibiting the generators. Is there a more abstract algebraic framework that would prove finite generation without computing the homology?

(4) Having described $H_*(\mathrm{config}(n, w))$, but not at all equivariantly, we can ask about its S_n–action by permuting the disk labels. In its decomposition into irreducible representations of S_n, how do the multiplicities grow in n? For finitely generated twisted noncommutative algebras in general, by what

patterns can the multiplicities grow? In particular, can one use presentations by generators and relations as in Section 8 to recover this information?

Appendix   Computer calculations for small *n*

In [Alpert et al. 2021], the Betti numbers of config(n, w) were computed for $n \leq 8$ using off-the-shelf software for computing persistent homology. This involved running the software on the complex cell(8), which has over 5 million cells. In general, cell(n) has $2^{n-1}n!$ cells.

We wrote a Python script that harnesses Theorems B and C to compute the persistence diagram of $H_*(\text{cell}(n, *))$. Although the runtime still grows as $n!$, this is faster than the above method by an exponential factor; for $n = 12$, the script ran on a laptop in less than 90 minutes. Figures 13, 14, 15 and 16 show a graphical representation of the resulting persistence diagram. We would like to thank Matthew Kahle for making the first version of this series of figures.

Figure 13: The persistent homology for the configuration space of 12 disks of unit diameter in a strip of width w. The thickness of a bar in the barcode is proportional to the logarithm of the multiplicity, and the exact multiplicity is in the rightmost column.
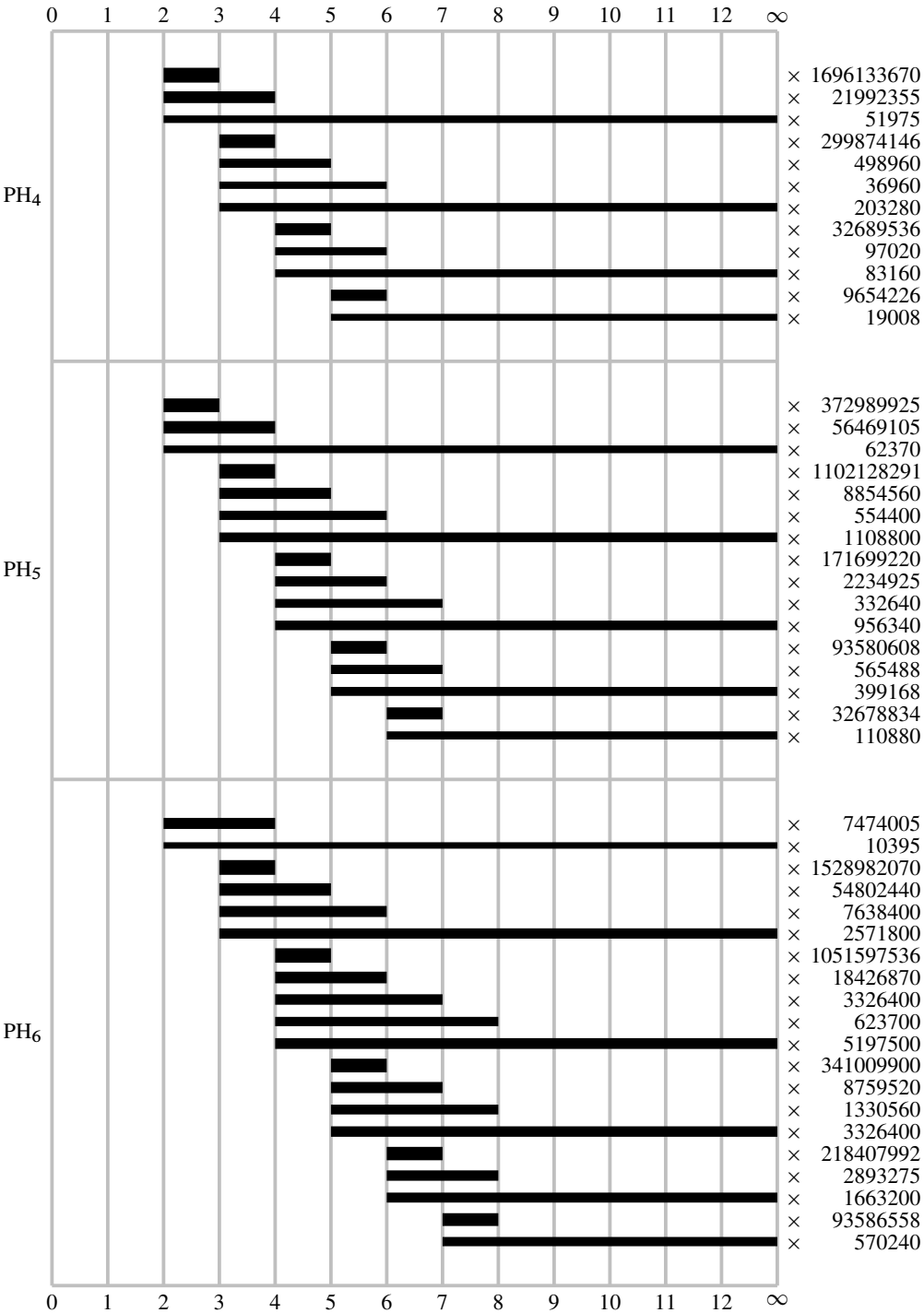
Figure 14: The persistent homology for the configuration space of 12 disks of unit diameter in a strip of width w. The thickness of a bar in the barcode is proportional to the logarithm of the multiplicity, and the exact multiplicity is in the rightmost column.
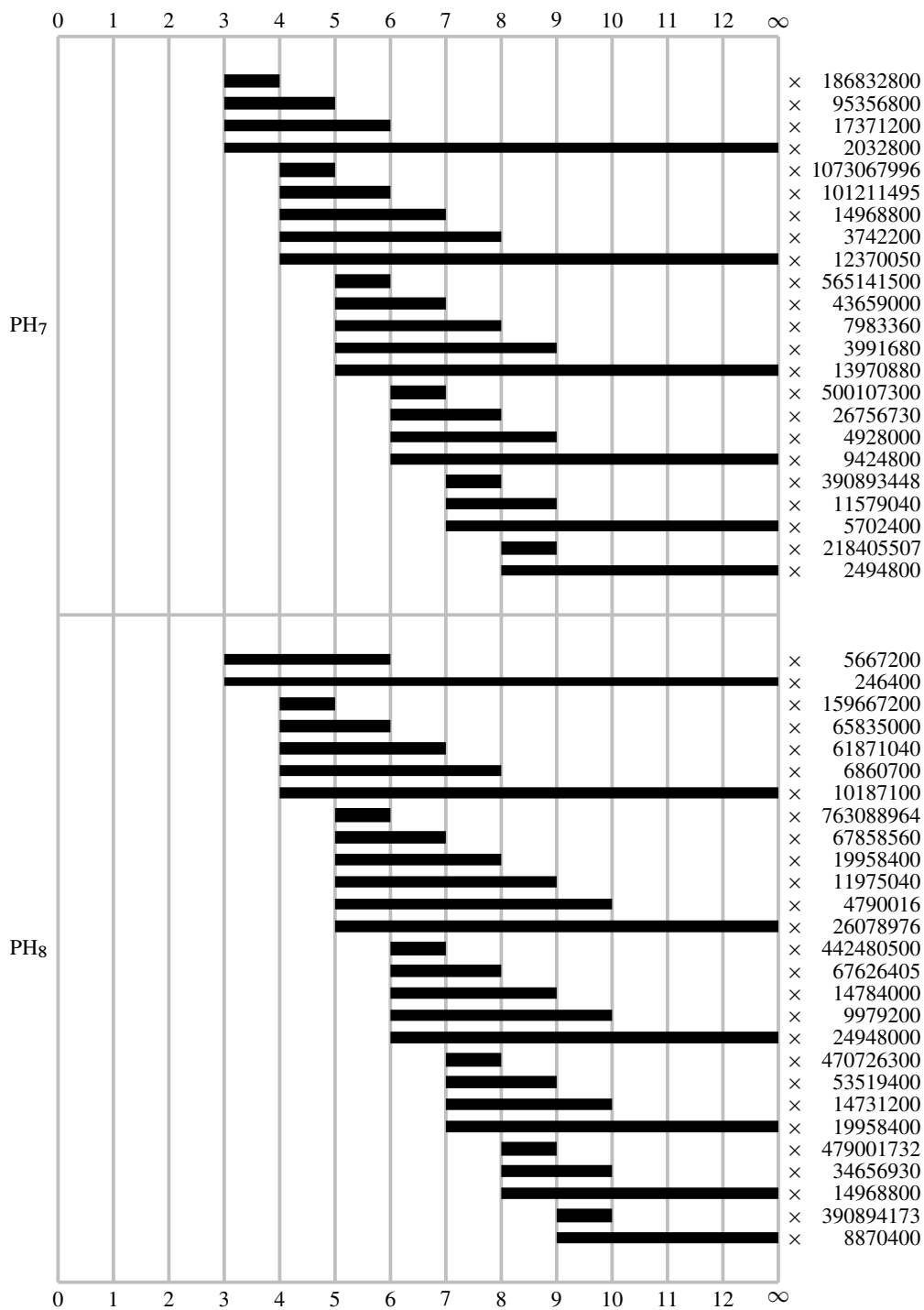
Figure 15: The persistent homology for the configuration space of 12 disks of unit diameter in a strip of width w. The thickness of a bar in the barcode is proportional to the logarithm of the multiplicity, and the exact multiplicity is in the rightmost column.
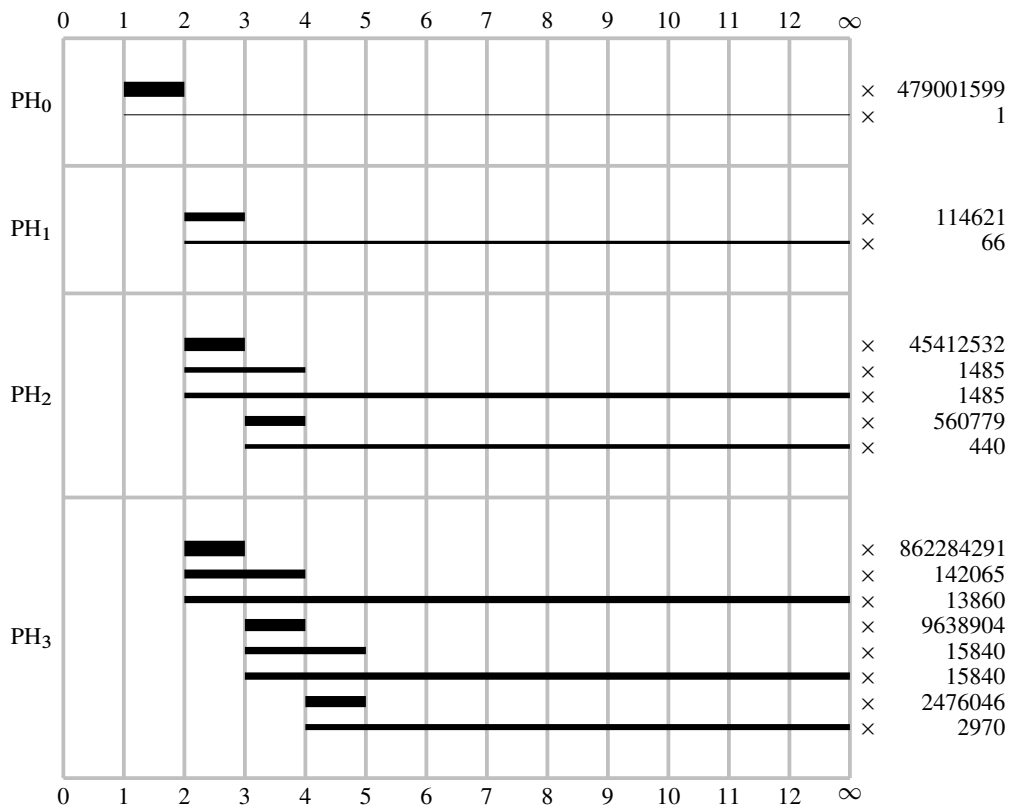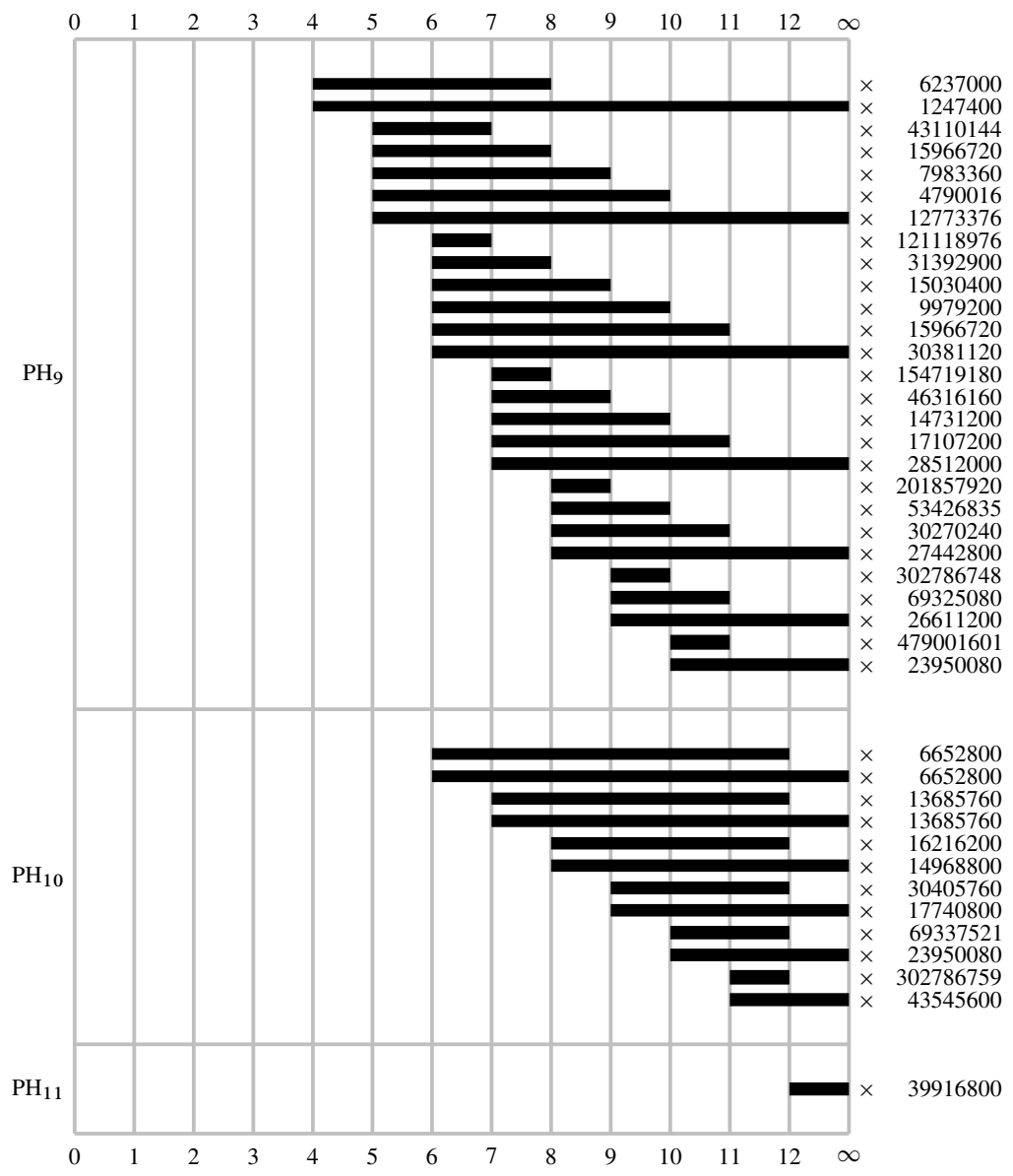
Figure 16: The persistent homology for the configuration space of 12 disks of unit diameter in a strip of width w. The thickness of a bar in the barcode is proportional to the logarithm of the multiplicity, and the exact multiplicity is in the rightmost column.

References

[Alpert 2020]   **H Alpert**, *Discrete configuration spaces of squares and hexagons*, J. Appl. Comput. Topol. 4 (2020) 263–280 MR Zbl

[Alpert et al. 2021]   **H Alpert**, **M Kahle**, **R MacPherson**, *Configuration spaces of disks in an infinite strip*, J. Appl. Comput. Topol. 5 (2021) 357–390 MR Zbl

[Alpert et al. 2023]   **H Alpert**, **U Bauer**, **M Kahle**, **R MacPherson**, **K Spendlove**, *Homology of configuration spaces of hard squares in a rectangle*, Algebr. Geom. Topol. 23 (2023) 2593–2626   MR Zbl

[Arnold 1969]   **V I Arnold**, *The cohomology ring of the colored braid group*, Mat. Zametki 5 (1969) 227–231   MR Zbl   In Russian; translated in Math. Notes 5 (1969) 138–140

[Baryshnikov et al. 2014]   **Y Baryshnikov**, **P Bubenik**, **M Kahle**, *Min-type Morse theory for configuration spaces of hard spheres*, Int. Math. Res. Not. 2014 (2014) 2577–2592   MR Zbl

[Bauer 2021]   **U Bauer**, *Ripser: efficient computation of Vietoris–Rips persistence barcodes*, J. Appl. Comput. Topol. 5 (2021) 391–423   MR Zbl

[Björner 2008]   **A Björner**, *Random walks, arrangements, cell complexes, greedoids, and self-organizing libraries*, from "Building bridges" (M Grötschel, G O H Katona, editors), Bolyai Soc. Math. Stud. 19, Springer (2008) 165–203   MR Zbl

[Björner and Welker 1995]   **A Björner**, **V Welker**, *The homology of "k–equal" manifolds and related partition lattices*, Adv. Math. 110 (1995) 277–313   MR Zbl

[Blagojević and Ziegler 2014]   **P V M Blagojević**, **G M Ziegler**, *Convex equipartitions via equivariant obstruction theory*, Israel J. Math. 200 (2014) 49–77   MR Zbl

[Carlsson et al. 2012]   **G Carlsson**, **J Gorham**, **M Kahle**, **J Mason**, *Computational topology for configuration spaces of hard disks*, Phys. Rev. E 85 (2012) art. id. 011303   Zbl

[Cohen 1976]   **F R Cohen**, *The homology of C_{n+1}-spaces, $n \geq 0$*, from "The homology of iterated loop spaces" (F R Cohen, T J Lada, J P May, editors), Lecture Notes in Math. 533, Springer (1976) 207–351   MR Zbl

[Diaconis 2009]   **P Diaconis**, *The Markov chain Monte Carlo revolution*, Bull. Amer. Math. Soc. 46 (2009) 179–205   MR Zbl

[Edelsbrunner and Harer 2008]   **H Edelsbrunner**, **J Harer**, *Persistent homology: a survey*, from "Surveys on discrete and computational geometry" (J E Goodman, J Pach, R Pollack, editors), Contemp. Math. 453, Amer. Math. Soc., Providence, RI (2008) 257–282   MR Zbl

[Farber 2008]   **M Farber**, *Invitation to topological robotics*, Eur. Math. Soc., Zürich (2008)   MR Zbl

[Forman 2002]   **R Forman**, *A user's guide to discrete Morse theory*, Sém. Lothar. Combin. 48 (2002) art. id. B48c   MR Zbl

[Kahle 2012]   **M Kahle**, *Sparse locally-jammed disk packings*, Ann. Comb. 16 (2012) 773–780   MR Zbl

[Mac Lane 1998]   **S Mac Lane**, *Categories for the working mathematician*, 2nd edition, Graduate Texts in Math. 5, Springer (1998)   MR Zbl

[Miller and Wilson 2019]   **J Miller**, **J C H Wilson**, *Higher-order representation stability and ordered configuration spaces of manifolds*, Geom. Topol. 23 (2019) 2519–2591   MR Zbl

[Ramos 2017]   **E Ramos**, *Generalized representation stability and FI_d–modules*, Proc. Amer. Math. Soc. 145 (2017) 4647–4660   MR Zbl

[Ramos 2019]   **E Ramos**, *Configuration spaces of graphs with certain permitted collisions*, Discrete Comput. Geom. 62 (2019) 912–944   MR Zbl

[Riehl 2017]   **E Riehl**, *Category theory in context*, Dover, Mineola, NY (2017)   Zbl

[Sam and Snowden 2012]   **S V Sam**, **A Snowden**, *Introduction to twisted commutative algebras*, preprint (2012)   arXiv 1209.5122

[Sam and Snowden 2017] **S V Sam**, **A Snowden**, *Gröbner methods for representations of combinatorial categories*, J. Amer. Math. Soc. 30 (2017) 159–203 MR Zbl

[Sinha 2013] **D P Sinha**, *The (non-equivariant) homology of the little disks operad*, from "OPERADS 2009" (J-L Loday, B Vallette, editors), Sémin. Congr. 26, Soc. Math. France, Paris (2013) 253–279 MR Zbl

[Ziegler 1995] **G M Ziegler**, *Lectures on polytopes*, Graduate Texts in Math. 152, Springer (1995) MR Zbl

[Zomorodian and Carlsson 2005] **A Zomorodian**, **G Carlsson**, *Computing persistent homology*, Discrete Comput. Geom. 33 (2005) 249–274 MR Zbl

Department of Mathematics and Statistics, Auburn University
Auburn, AL, United States

Department of Mathematics, University of California, Santa Barbara
Santa Barbara, CA, United States

hcalpert@auburn.edu, manin@math.ucsb.edu

http://webhome.auburn.edu/~hca0013/, http://web.math.ucsb.edu/~manin/

Closed geodesics with prescribed intersection numbers

Yann Chaubet

Let (Σ, g) be a closed oriented negatively curved surface, and fix a simple closed geodesic $\gamma_\star$. We give the asymptotic growth as $L \to +\infty$ of the number of primitive closed geodesics of length less than L intersecting $\gamma_\star$ exactly n times, where n is fixed positive integer. This is done by introducing a dynamical scattering operator associated to the surface with boundary obtained by cutting Σ along $\gamma_\star$ and by using the theory of Pollicott–Ruelle resonances for open systems.

1 Introduction

Let (Σ, g) be a closed oriented connected negatively curved Riemannian surface, and denote by $\mathcal{P}$ the set of its oriented primitive closed geodesics. For $L > 0$ define

$$N(L) = \#\{\gamma \in \mathcal{P} : \ell(\gamma) \leqslant L\},$$

where, for $\gamma \in \mathcal{P}$, we denote by $\ell(\gamma)$ its length. Then a classical result obtained by Margulis [31] states that

$$N(L) \sim \frac{e^{hL}}{hL} \quad \text{as } L \to \infty,$$

where $h > 0$ is the topological entropy of the geodesic flow of (Σ, g).

Our purpose here is to provide a similar asymptotic result for closed geodesics satisfying certain intersection constraints. Namely, let $\gamma_\star$ be a simple closed geodesic of (Σ, g). For any $\gamma \in \mathcal{P}$, we denote by $i(\gamma, \gamma_\star)$ the geometric intersection number between γ and $\gamma_\star$ (see Section 2.1), and we set

$$N(n, L) = \#\{\gamma \in \mathcal{P} : \ell(\gamma) \leqslant L \text{ and } i(\gamma, \gamma_\star) = n\}.$$

We first state a result assuming $\gamma_\star$ is not separating, in the sense that $\Sigma \setminus \gamma_\star$ is connected.

Theorem 1 *Assume that $\gamma_\star$ is not separating. Then there are $c_\star > 0$ and $h_\star \in {]}0, h{[}$ such that, for any $n \geqslant 1$,*

$$(1\text{-}1) \qquad\qquad N(n, L) \sim \frac{(c_\star L)^n}{n!} \frac{e^{h_\star L}}{h_\star L} \quad \text{as } L \to \infty.$$

The number $h_\star$ in the above statement is the topological entropy of the geodesic flow (φ_t) of (Σ, g) when restricted to the trapped set

$$K_\star = \overline{\{(x, v) \in S\Sigma : \pi(\varphi_t(x, v)) \in \Sigma \setminus \gamma_\star \text{ for } t \in \mathbb{R}\}},$$

where the closure is taken in $S\Sigma$ and $\pi \colon S\Sigma \to \Sigma$ is the natural projection. Also, we provide in Section 7 a description of the constant $c_\star$ in terms of the Pollicott–Ruelle resonant states of the geodesic flow of the compact surface with boundary $\Sigma_\star$ obtained by cutting Σ along $\gamma_\star$.

By using a classical large deviation result by Kifer [25] and Bonahon's intersection form [6], one is able to show that a typical closed geodesic γ satisfies $i(\gamma, \gamma_\star) \approx I_\star \ell(\gamma)$ for some $I_\star > 0$ not depending on γ (see Proposition 8.1 for a precise statement). In particular, Theorem 1 is a statement about very uncommon closed geodesics.

The asymptotics (1-1) for $n = 0$ is well known and follows from the work of Dal'bo [12] and from the growth rate of periodic orbits of axiom A flows obtained by Parry and Pollicott [35] (see Section 2.5). However, to the best of our knowledge, the result is new for $n > 0$. Note that it would be tempting to sum the right-hand side of (1-1) over n in order to recover the asymptotic growth of $N(L)$ — for example, one could hope that $h_\star + c_\star = h$ — but if L is fixed, the left-hand side of (1-1) vanishes whenever n is large enough, and it is very unlikely that such an equality holds.

If $\gamma_\star$ is separating then $i(\gamma, \gamma_\star)$ is even, and we have the following result:

Theorem 2 *Suppose that $\gamma_\star$ separates Σ in two surfaces, Σ_1 and Σ_2. Let $h_j \in\,]0, h[$ denote the entropy of the open system $(\Sigma_j, g|_{\Sigma_j})$ and set $h_\star = \max(h_1, h_2)$. Then there is $c_\star > 0$ such that, for each $n \geqslant 1$, as $L \to +\infty$,*

$$
N(2n, L) \sim
\begin{cases}
\dfrac{(c_\star L)^n}{n!} \dfrac{e^{h_\star L}}{h_\star L} & \text{if } h_1 \neq h_2, \\[2ex]
2\dfrac{(c_\star L^2)^n}{(2n)!} \dfrac{e^{h_\star L}}{h_\star L} & \text{if } h_1 = h_2.
\end{cases}
$$

As before, the number h_j is defined as the topological entropy of the geodesic flow restricted to the trapped set

$$
K_j = \overline{\{(x, v) \in S\Sigma : \pi(\varphi_t(x, v)) \in \Sigma_j \setminus \gamma_\star \text{ for } t \in \mathbb{R}\}},
$$

where the closure is taken in $S\Sigma$.

We also have an equidistribution result, as follows. Set

$$
\partial_\star = \{(x, v) \in S\Sigma : x \in \gamma_\star\} \quad \text{and} \quad \Gamma = S\gamma_\star \cup \{z \in \partial_\star : \varphi_t(z) \in S\Sigma \setminus \partial_\star \text{ for } t > 0\},
$$

where $S\gamma_\star = \{(x, v) \in \partial_\star : v \in T_x\gamma_\star\}$. We define the scattering map $S \colon \partial_\star \setminus \Gamma \to \partial_\star$ by

$$
S(z) = \varphi_{\ell(z)}(z), \quad \ell(z) = \inf\{t > 0 : \varphi_t(z) \in \partial_\star\} \quad \text{for } z \in \partial_\star \setminus \Gamma.
$$

For any $n \in \mathbb{N}_{\geqslant 1}$ we set

$$
\Gamma_n = \partial_\star \setminus \{z \in \partial_\star \setminus \Gamma : S^k(z) \in \partial_\star \setminus \Gamma \text{ for } k = 1, \ldots, n-1\},
$$

which is a closed set of Lebesgue measure zero, and

$$
\ell_n(z) = \ell(z) + \cdots + \ell(S^{n-1}(z)) \quad \text{for } z \in \partial_\star \setminus \Gamma_n.
$$

Theorem 3 *Assume that $\gamma_\star$ is not separating and let $n \geqslant 1$. For any $f \in C^\infty(\partial_\star)$, the limit*

$$\lim_{L \to +\infty} \frac{1}{N(n, L)} \sum_{\substack{\gamma \in \mathcal{P} \\ i(\gamma, \gamma_\star)=n}} \frac{1}{\#I_\star(\gamma)} \sum_{z \in I_\star(\gamma)} f(z)$$

exists, where, for any $\gamma \in \mathcal{P}$, the set $I_\star(\gamma) = \{(x, v) \in S\gamma : x \in \gamma_\star\}$ consists of the incidence vectors of γ along $\gamma_\star$. This formula defines a probability measure μ_n on $\partial_\star$, whose support is contained in Γ_n.

Of course, a similar statement holds even if $\gamma_\star$ is separating, though we will not explicitly state it here. As for $c_\star$, we will provide a full description of μ_n in terms of the Pollicott–Ruelle resonant states of the geodesic flow of $(\Sigma_\star, g)$ for the resonance $h_\star$ in Section 7. Here, as before, $\Sigma_\star$ is the compact surface with boundary obtained by cutting Σ along $\gamma_\star$ (see Section 2.5).

Strategy of proof

A key ingredient used in the proof of Theorems 1, 2 and 3 is the scattering operator $\mathcal{S}(s): C^\infty(\partial_\star) \to C^\infty(\partial_\star \setminus \Gamma)$, which is defined by

$$\mathcal{S}(s)f(z) = f(S(z))e^{-s\ell(z)} \quad \text{for } z \in \partial_\star \setminus \Gamma \text{ and } s \in \mathbb{C}.$$

As a first step (which is of independent interest; see the corollary on page 714), we prove that, for any $\chi \in C_c^\infty(\partial_\star \setminus S\gamma_\star)$, the family $s \mapsto \chi\mathcal{S}(s)\chi$ extends to a meromorphic family of operators $\mathcal{S}(s): C^\infty(\partial_\star) \to \mathcal{D}'(\partial_\star)$ on the whole complex plane (here $\mathcal{D}'(\partial_\star)$ denotes the space of distributions on $\partial_\star$), whose poles are contained in the set of Pollicott–Ruelle resonances of the geodesic flow of the surface with boundary $(\Sigma_\star, g)$; see Section 2.6 for the definition of those resonances. In this context, the existence of such resonances follows from the work of Dyatlov and Guillarmou [15], and we relate $\mathcal{S}(s)$ with the resolvent of the geodesic flow (see Proposition 3.2). By using the microlocal structure of the resolvent of the geodesic flow provided by [15], we are moreover able to prove that the composition $(\chi\mathcal{S}(s)\chi)^n$ is well defined for any $n \geqslant 1$, as well as its superflat trace (meaning that we also look at the action of $\mathcal{S}(s)$ on differential forms, see Section 3.4), which reads

$$(1\text{-}2) \qquad \operatorname{tr}_s^\flat[(\chi\mathcal{S}(s)\chi)^n] = n \sum_{i(\gamma, \gamma_\star)=n} \frac{\ell^\#(\gamma)}{\ell(\gamma)} e^{-s\ell(\gamma)} \prod_{z \in I_\star(\gamma)} \chi^2(z),$$

where the products runs over all closed geodesics (not necessarily primitive) γ with $i(\gamma, \gamma_\star) = n$, and $\ell^\#(\gamma)$ is the primitive length of γ. This formula will be obtained by using the Atiyah–Bott trace formula [3], though our scattering map S has singularities that we have to deal with. Furthermore, using a priori bounds on the growth of $N(n, L)$ (obtained in Section 4 by purely geometric techniques coming from the theory of CAT(-1) spaces), we prove that $s \mapsto \operatorname{tr}_s^\flat[(\chi\mathcal{S}(s)\chi)^n]$ has a pole of order n at $s = h_\star$ provided that χ has enough support. For this step, we crucially use the fact that the asymptotics for $N(0, L)$ is already known, although we could recover it by using the modern techniques introduced in [15] without going

through the scattering maps. Finally, letting the support of $1 - \chi$ be very close to $S\gamma_\star$, and estimating the growth of geodesics having n intersections with $\gamma_\star$ with at least one small angle, we are able to derive Theorems 1 and 2 from a classical Tauberian theorem of Delange [14].

Related works

As mentioned before, the case $n = 0$ follows from work of Parry and Pollicott [35] which is based on important contributions of Bowen [9; 10], as the geodesic flow on $(\Sigma_\star, g)$ can be seen as an axiom A flow; see Lemma 2.5 below and [15, Section 6.1]. For counting results on noncompact Riemann surfaces, see also the works of Sarnak [43], Guillopé [21] or Lalley [27]. We refer to the work of Paulin, Pollicott and Schapira [37] for counting results in more general settings.

We also mention a result by Pollicott [39] which says that, if (Σ, g) is of constant curvature -1 and if $\gamma_\star$ is not separating,

$$(1\text{-}3) \qquad \frac{1}{N(L)} \sum_{\substack{\gamma \in \mathcal{P} \\ \ell(\gamma) \leqslant L}} i(\gamma, \gamma_\star) \sim I_\star L$$

for some $I_\star > 0$. Roughly speaking, this means that the average intersection number between $\gamma_\star$ and closed geodesics of length not greater than L is about $I_\star L$. We will show that this result also holds in our context (see Section 8.2).

Lalley [26], Pollicott [40] and Anantharaman [1] investigated the asymptotic growth of the number of closed geodesics satisfying some homological constraints (see also Phillips and Sarnak [38] and Katsuda and Sunada [24] for the constant curvature case). They showed that, for any homology class $\xi \in H_1(\Sigma, \mathbb{Z})$,

$$\#\{\gamma \in \mathcal{P} : \ell(\gamma) \leqslant L \text{ and } [\gamma] = \xi\} \sim C e^{hL} / L^{g+1}$$

for some $C > 0$ independent of ξ, where g is the genus of Σ and $h > 0$ is the topological entropy of the geodesic flow of (Σ, g). Such asymptotics are obtained by studying L–functions associated to some characters of $H_1(\Sigma, \mathbb{Z})$. However, our problem is very different in nature; indeed, fixing a constraint in homology boils down to fixing *algebraic* intersection numbers, whereas here we are interested in *geometric* intersection numbers. In particular, L–functions are not well suited for this situation.

In the context of hyperbolic surfaces (ie surfaces with constant negative curvature -1), Mirzakhani [32; 33] computed the asymptotic growth of closed geodesics with prescribed self-intersection numbers. Namely, for any $k \in \mathbb{N}$,

$$\#\{\gamma \in \mathcal{P} : \ell(\gamma) \leqslant L \text{ and } i(\gamma, \gamma) = k\} \sim c_k L^{6(g-1)},$$

where $i(\gamma, \gamma)$ denotes the self-intersection number of γ; see also Erlandsson and Souto [17].

Note that our scattering map S defined above shares some similarities with the Sinai billiard map [44]. Similarly to the map S, which is not defined on the singularity set Γ, the billiard map is not continuous near some singular set consisting in grazing trajectories. In particular, it is plausible that recent functional analytic techniques developed by Baladi, Demers and Liverani [5] (see also Baladi and Demers [4]), as

the Sinai billiard map could be used to define an intrinsic spectrum of resonances for the transfer operator associated to S (without going through the resolvent of the geodesic flow of $S\Sigma_\star$).

We finally mention that the techniques presented herein allow one to obtain the asymptotic growth of closed geodesics for which *several* intersection numbers (with a family pairwise disjoint simple closed curves) are prescribed. However, such an extension requires more work, and for simplicity we will focus here on the case where we are given only one simple geodesic. The aforementioned generalization will be the subject of subsequent work.

Organization of the paper

The paper is organized as follows. In Section 2 we introduce some geometric and dynamical tools. In Section 3 we introduce the dynamical scattering operator, which is a central object in this paper, and we compute its flat trace. In Section 4 we prove a priori bounds on $N(n, L)$. In Section 5 we use a Tauberian argument to estimate certain quantities. In Section 6 we prove Theorems 1 and 2. In Section 7 we prove Theorem 3. Finally, in Section 8 we show that a typical closed geodesic γ satisfies $i(\gamma, \gamma_\star) \approx I_\star \ell(\gamma)$ for some $I_\star > 0$.

Acknowledgements

2 Geometric preliminaries

We recall here some classical geometric and dynamical notions, and introduce the Pollicott–Ruelle resonances that will arise in our situation. Throughout the article, (Σ, g) will denote a closed connected oriented Riemannian surface of negative curvature.

2.1 Geometric intersection numbers

For any two loops $\alpha, \beta \colon \mathbb{R}/\mathbb{Z} \to \Sigma$, the *geometric intersection number* between α and β is defined by

$$i(\alpha, \beta) = \inf_{\alpha' \sim \alpha, \beta' \sim \beta} |\alpha \cap \beta|,$$

where the infimum runs over all loops α' and β' freely homotopic to α and β, respectively, and

$$|\alpha \cap \beta| = \{(\tau, \tau') \in (\mathbb{R}/\mathbb{Z})^2 : \alpha(\tau) = \beta(\tau')\}.$$

It is well known that, in every nontrivial free homotopy class of loops c, there is a unique oriented closed geodesic $\gamma_c \in c$ which minimizes the length among curves in c. In fact, closed geodesics also minimize intersection numbers, as follows:

Lemma 2.1 *Let γ_1 and γ_2 be any two nontrivial oriented closed geodesics, and assume that γ_1 (resp. γ_2) is not freely homotopic to a power of γ_2 (resp. γ_1). Then*

$$i(\gamma_1, \gamma_2) = |\gamma_1 \cap \gamma_2|.$$

The above result is rather classical, but for the reader's convenience we provide a proof in Appendix A.

2.2 Structural equations

Here we recall some classical facts from [45, Section 7.2] about geometry of surfaces. Denote by $M = S\Sigma = \{(x, v) \in T\Sigma : \|v\|_g = 1\}$ the unit tangent bundle of Σ, and by X the geodesic vector field on M, that is, the generator of the geodesic flow $\varphi = (\varphi_t)_{t \in \mathbb{R}}$ of (Σ, g), acting on M. The Liouville one-form α on M is defined by

$$\langle \alpha(z), \eta \rangle = \langle d_{(x,v)}\pi(\eta), v \rangle \quad \text{for } z = (x, v) \in M \text{ and } \eta \in T_{(x,v)}M,$$

where $\pi \colon M \to \Sigma$ is the natural projection. Then α is a contact form (that is, $\alpha \wedge d\alpha$ is a volume form on M) and it turns out that X is the Reeb vector field associated to α, meaning that

$$\iota_X \alpha = 1 \quad \text{and} \quad \iota_X d\alpha = 0,$$

where ι denotes the interior product.

We also set $\beta = R_{\pi/2}^* \alpha$, where, for $\theta \in \mathbb{R}$, $R_\theta \colon M \to M$ is the rotation of angle θ in the fibers. Finally we denote by ψ the connection one-form, defined as the unique one-form on M satisfying

$$\iota_V \psi = 1, \quad d\alpha = \psi \wedge \beta \quad \text{and} \quad d\beta = -\psi \wedge \alpha,$$

where V is the vertical vector field, that is, the vector field generating $(R_\theta)_{\theta \in \mathbb{R}}$. Then (α, β, ψ) is a global frame of T^*M, and we denote by H the unique vector field on M such that (X, H, V) is the dual frame of (α, β, ψ). We then have the commutation relations

$$[V, X] = H, \quad [V, H] = -X \quad \text{and} \quad [X, H] = (\kappa \circ \pi)V,$$

where κ is the Gauss curvature of (Σ, g).

2.3 The Anosov property

It is known, by the work of Anosov [2], that the flow (φ_t) is hyperbolic. That is, for any $z \in M$ there is a $d\varphi_t$–invariant splitting

$$T_z M = \mathbb{R}X(z) \oplus E_s(z) \oplus E_u(z)$$

which depends continuously on z, and has the property that, for any norm $\|\cdot\|$ on TM, there exist $C, \nu > 0$ such that

$$\|\mathrm{d}\varphi_t(z)v\| \leqslant Ce^{-\nu t}\|v\| \quad \text{for } v \in E_s(z), \ t \geqslant 0 \text{ and } z \in M,$$

and

$$\|\mathrm{d}\varphi_{-t}(z)v\| \leqslant Ce^{-\nu t}\|v\| \quad \text{for } v \in E_u(z), \ t \geqslant 0 \text{ and } z \in M.$$

In fact, $E_s(z) \oplus E_u(z) = \ker \alpha(z)$ and there exist two continuous functions $r_\pm \colon M \to \mathbb{R}$ such that $\pm r_\pm > 0$ and

$$E_s(z) = \mathbb{R}(H(z) + r_- V(z)) \quad \text{and} \quad E_u(z) = \mathbb{R}(H(z) + r_+ V(z)) \quad \text{for } z \in M.$$

Moreover, the functions $r_\pm$ are differentiable along the flow direction, and they satisfy the Riccati equation

$$\pm X r_\pm + r_\pm^2 + \kappa \circ \pi = 0,$$

where κ is the curvature of Σ.

We will denote by $T^*M = E_0^* \oplus E_s^* \oplus E_u^*$ the splitting defined by

$$E_0^*(E_u \oplus E_s) = 0, \quad E_s^*(E_s \oplus E_0) = 0, \quad E_u^*(E_u \oplus E_0) = 0.$$

(Here the bundle $\mathbb{R}X$ is denoted by E_0.) Then we have $E_0^* = \mathbb{R}\alpha$ and

(2-1) $$E_s^* = \mathbb{R}(r_- \beta - \psi), \quad E_u^* = \mathbb{R}(r_+ \beta - \psi).$$

Note that this decomposition does not coincide with the usual dual decomposition, but it is motivated by the fact that covectors in E_s^* (resp. E_u^*) are exponentially contracted in the future (resp. in the past) by the symplectic lift Φ_t of φ_t, which is defined by

(2-2) $$\Phi_t(z, \xi) = (\varphi_t(z), \mathrm{d}\varphi_t(z)^{-\top} \cdot \xi) \quad \text{for } (z, \xi) \in T^*M \text{ and } t \in \mathbb{R},$$

where $^{-\top}$ denotes the inverse transpose. We have the following lemma:

Lemma 2.2 [13, Section 3.2] *If $t \neq 0$, we have $\iota_V \Phi_t(\beta) \neq 0$ and $\iota_H \Phi_t(\psi) \neq 0$.*

2.4 A nice system of coordinates

In what follows, we write

$$\partial_\star = \{(x, v) \in M : x \in \gamma_\star\} = S\Sigma|_{\gamma_\star}.$$

Lemma 2.3 *There exists a tubular neighborhood U of $\partial_\star$ in M, and coordinates (τ, ρ, θ) on U with*

$$U \simeq (\mathbb{R}/\ell_\star \mathbb{Z})_\tau \times (-\delta, \delta)_\rho \times (\mathbb{R}/2\pi\mathbb{Z})_\theta,$$

where $\ell_\star$ is the length of $\gamma_\star$, and such that

$$|\rho(z)| = \mathrm{dist}_g(\pi(z), \gamma_\star) \quad \text{and} \quad S_z\Sigma = \{(\tau(z), \rho(z), \theta) : \theta \in \mathbb{R}/2\pi\mathbb{Z}\} \quad \text{for } z \in U.$$

Moreover, in these coordinates, on $\{\rho = 0\}$,

$$X = \cos(\theta)\partial_\tau + \sin(\theta)\partial_\rho, \quad H = -\sin(\theta)\partial_\tau + \cos(\theta)\partial_\rho, \quad V = \partial_\theta,$$

and

$$\alpha = \cos(\theta)\,\mathrm{d}\tau + \sin(\theta)\,\mathrm{d}\rho, \quad \beta = -\sin(\theta)\,\mathrm{d}\tau + \cos(\theta)\,\mathrm{d}\rho, \quad \psi = \mathrm{d}\theta.$$

Proof For $\tau \in \mathbb{R}/\ell_\star\mathbb{Z}$ we set $(x_\tau, v_\tau) = \varphi_\tau(\gamma_\star(0), \dot{\gamma}_\star(0))$. We now define, for $\delta > 0$ small enough,

$$\Psi(\tau, \rho, \theta) = R_{\theta - \pi/2}\varphi_\rho(x_\tau, \nu(x_\tau)) \quad \text{for } (\tau, \rho, \theta) \in \mathbb{R}/\ell_\star\mathbb{Z} \times (-\delta, \delta) \times \mathbb{R}/2\pi\mathbb{Z},$$

where $R_\eta: S\Sigma \to S\Sigma$ is the rotation of angle η and $\nu(x_\tau) = R_{\pi/2}v_\tau$. Then $\mathrm{d}\Psi(\tau, 0, \theta)$ is injective for any τ and θ. Indeed, $\partial_\tau(\pi \circ \Psi)(\tau, 0, \theta) = v_\tau$ and $\partial_\rho(\pi \circ \Psi)(\tau, 0, \theta) = \nu(x_\tau)$. Thus $\mathrm{d}\Psi(\tau, 0, \theta): \mathbb{R}\partial_\tau \oplus \mathbb{R}\partial_\rho \to T\Sigma$ is injective. Moreover, $\partial_\theta(\pi \circ \Psi)(\tau, 0, \theta) = 0$ and $\partial_\theta\Psi(\tau, 0, \theta) = V(\Psi(\tau, 0, \theta)) \neq 0$. Thus $\mathrm{d}\Psi(\tau, 0, \theta)$ is injective for any τ and θ, and furthermore, if $\delta > 0$ is small enough, $\Psi: U \to M$ is an immersion. In particular, since $(\tau, \theta) \mapsto \Psi(\tau, 0, \theta)$ is clearly injective, we obtain that $\Psi|_U$ is a diffeomorphism onto its image provided that δ is chosen small enough.

Because $V = \partial_\theta$ and $\iota_V\alpha = \iota_V\beta = 0$, we may write $\alpha(\tau, 0, \theta) = a(\tau, \theta)\,\mathrm{d}\tau + b(\tau, \theta)\,\mathrm{d}\rho$ and $\beta(\tau, 0, \theta) = a'(\tau, \theta)\,\mathrm{d}\tau + b'(\tau, \theta)\,\mathrm{d}\rho$ for some smooth functions a, a', b and b'. Now, since $\mathrm{d}\alpha = \psi \wedge \beta$, we obtain $\mathcal{L}_V\alpha = \iota_V\,\mathrm{d}\alpha = \beta$, and similarly $\mathcal{L}_V\beta = -\alpha$. Thus, $a' = \partial_\theta a$, $b' = \partial_\theta b$ and

$$\partial_\theta^2 a + a = 0, \quad \partial_\theta^2 b + b = 0.$$

In consequence, $a(\tau, \theta) = a_1(\tau)\cos\theta + a_2(\tau)\sin\theta$ and $b(\tau, \theta) = b_1(\tau)\cos\theta + b_2(\tau)\sin\theta$ for some smooth functions a_1, a_2, b_1 and b_2. Moreover, by definition of the coordinates (τ, ρ, θ), one has

(2-3) $$X(\tau, 0, 0) = \partial_\tau \quad \text{and} \quad X(\tau, 0, \tfrac{1}{2}\pi) = \partial_\rho.$$

Therefore $a_1 = b_2 = 1$ and $a_2 = b_1 = 0$. We thus get the desired formulae for α and β. Now, writing $\psi = a''\,\mathrm{d}\tau + b''\,\mathrm{d}\rho + \mathrm{d}\theta$ and using $\mathcal{L}_V\psi = 0$, we obtain $\partial_\theta a'' = \partial_\theta b'' = 0$. As $\iota_X\psi = 0$ we obtain $a'' = b'' = 0$ by (2-3). The formulae for X, H and V follow. $\qquad\square$

Remark 2.4 If $\tilde{\partial} = \{\rho = 0\}$, then, for any $z = (\tau, 0, \theta) \in \partial$,

$$T_z\tilde{\partial} = \mathbb{R}V(z) \oplus \mathbb{R}(\cos(\theta)X(z) - \sin(\theta)H(z)) \quad \text{and} \quad N_z^*\tilde{\partial} = \mathbb{R}(\sin(\theta)\alpha(z) + \cos(\theta)\beta(z)).$$

2.5 Cutting the surface along $\gamma_\star$

As mentioned in the introduction, we may see $\Sigma \setminus \gamma_\star$ as the interior of a compact surface $\Sigma_\star$ with boundary consisting of two copies of $\gamma_\star$. By gluing two copies of the annulus U obtained in the preceding subsection on each component of the boundary of $\Sigma_\star$, we construct a slightly larger surface $\Sigma_\delta \supset \Sigma_\star$ whose boundary is identified with the boundary of U (see Figure 1).

Lemma 2.5 *The surface Σ_δ has strictly convex boundary, in the sense that the second fundamental form of the boundary $\partial\Sigma_\delta$ with respect to its outward normal pointing vector is strictly negative.*

Proof In the coordinates (τ, ρ) given by Lemma 2.3, the metric g has the form

(2-4) $$\mathrm{d}\rho^2 + f(\tau, \rho)\,\mathrm{d}\tau^2$$

for some $f > 0$ satisfying $\partial_\rho f(\tau, 0) = 0$. Indeed, if ∇ is the Levi-Civita connection, one has

$$\frac{\mathrm{d}}{\mathrm{d}\rho}\langle \partial_\rho, \partial_\tau \rangle = \langle \nabla_{\partial_\rho}\partial_\rho, \partial_\tau \rangle + \langle \partial_\rho, \nabla_{\partial_\rho}\partial_\tau \rangle = \langle \partial_\rho, \nabla_{\partial_\tau}\partial_\rho \rangle = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\langle \partial_\rho, \partial_\rho \rangle = 0,$$
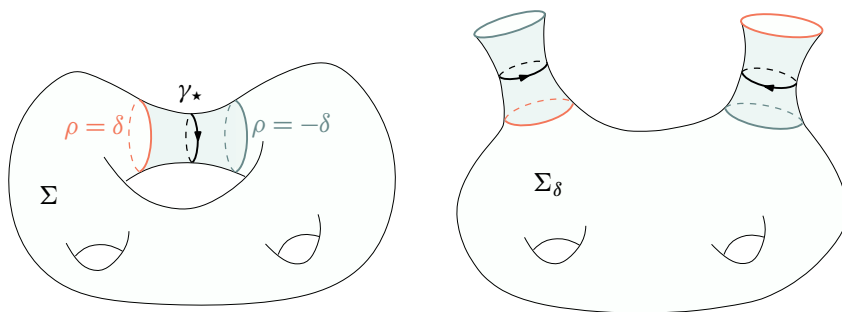
Figure 1: The surfaces Σ (on the left) and Σ_δ (on the right) in the case where $\gamma_\star$ is not separating. In Σ, the darker region corresponds to the neighborhood $\pi(U)$ of $\gamma_\star$.

since $\nabla_{\partial_\rho} \partial_\rho = 0$ (indeed, $\rho \mapsto (\tau, \rho)$ is a geodesic curve). Thus $\langle \partial_\tau, \partial_\rho \rangle = \langle \partial_\tau, \partial_\rho \rangle|_{\rho=0} = 0$. In particular, g has the form (2-4) with $f(\tau, \rho) = \langle \partial_\tau, \partial_\tau \rangle$, and we have $\partial_\rho f(\tau, 0) = \partial_\rho \langle \partial_\tau, \partial_\tau \rangle = 2 \partial_\tau \langle \partial_\rho, \partial_\tau \rangle|_{\rho=0} = 0$ (indeed, since $\tau \mapsto (\tau, 0)$ is a geodesic curve, $\nabla_{\partial_\tau} \partial_\tau = 0$ on $\{\rho = 0\}$). In those coordinates, the scalar curvature reads

$$\kappa(\tau, \rho) = \frac{-\partial_\rho^2 f(\tau, \rho)}{f(\tau, \rho)}.$$

As $\kappa < 0$, we get $\partial_\rho^2 f > 0$, which gives $\pm \partial_\rho f > 0$ on $\{\pm \rho > 0\}$. The second fundamental form of $\partial \Sigma_\delta$ with respect to ∂_ρ is defined by

$$\langle \nabla_{\partial_\tau} \partial_\tau, \partial_\rho \rangle = -\tfrac{1}{2} \partial_\rho f(\tau, \rho),$$

which concludes the proof, since ∂_ρ is outward pointing (resp. inward pointing) on $\{\rho = \delta\}$ (resp. $\{\rho = -\delta\}$). $\qquad \square$

Lemma 2.6 *In the coordinates given by Lemma 2.3,*

$$\pm X^2 \rho > 0 \quad on \ \{\pm \rho > 0\}.$$

Proof Since, in the coordinates (τ, ρ), the metric g has the form (2-4), the Christoffel symbols of g are given by

$$\Gamma_{\rho\rho}^\rho = \Gamma_{\tau\rho}^\rho = 0 \quad \text{and} \quad \Gamma_{\tau\tau}^\rho = -\tfrac{1}{2} \partial_\rho f.$$

In particular, if $t \mapsto (\tau(t), \rho(t))$ is a geodesic path,

$$\ddot{\rho}(t) - \tfrac{1}{2} \partial_\rho f(\tau(t), \rho(t)) = 0.$$

Because $\partial_\rho f(\tau, 0) = 0$ and $-\partial_\rho^2 f / f = \kappa < 0$, we obtain that $\pm \partial_\rho f > 0$ whenever $\pm \rho > 0$. $\qquad \square$

2.6 The resolvent of the geodesic flow for open systems

In what follows, we denote by $\Omega^\bullet(M_\delta)$ the set of differential forms on M_δ and by $\Omega_c^\bullet(M_\delta)$ the elements of $\Omega^\bullet(M_\delta)$ whose support is contained in the interior of M_δ. Here $M_\delta = S\Sigma_\delta$ is the unit tangent bundle

of Σ_δ. The set of currents on M_δ, denoted by $\mathcal{D}'^\bullet(M_\delta)$, is defined as the topological dual of $\Omega_c^\bullet(M_\delta)$. Note that we have an inclusion $\Omega^\bullet(M_\delta) \hookrightarrow \mathcal{D}'^\bullet(M_\delta)$ via the pairing

$$\langle u, v \rangle = \int_{M_\delta} u \wedge v \quad \text{for } u, v \in \Omega^\bullet(M_\delta).$$

The geodesic flow φ on M induces a flow on $M_\delta = S\Sigma_\delta$, which we still denote by φ. We set

$$\partial_\pm M_\delta = \{(x, v) \in \partial M_\delta : \pm\langle v, \nu_\delta(x) \rangle > 0\} \quad \text{and} \quad \partial_0 M_\delta = \{(x, v) \in \partial M_\delta : \pm\langle v, \nu_\delta(x) \rangle = 0\},$$

where $\nu_\delta(x)$ is the unit vector orthogonal to $\partial\Sigma_\delta$, based at x, and pointing outward. Next, define

$$\ell_{\pm,\delta}(z) = \inf\{t > 0 : \varphi_{\pm t}(z) \in \partial M_\delta\} \quad \text{for } z \in \text{int}(M_\delta) \cup \partial_\mp M_\delta,$$

and $\ell_{\pm,\delta}(z) = 0$ for $z \in \partial_\pm M_\delta \cup \partial_0 M_\delta$, where $\text{int}(M_\delta)$ denotes the interior of M_δ. The numbers $\ell_{\pm,\delta}(z)$ are the first exit times of z in the future and in the past. We also set

$$\Gamma_{\pm,\delta} = \{z \in M_\delta : \ell_\mp(z) = +\infty\} \quad \text{and} \quad K_\delta = \Gamma_\delta^+ \cap \Gamma_\delta^-,$$

and we define the operators $R_{\pm,\delta}(s)$ by

$$(2\text{-}5) \qquad R_{\pm,\delta}(s)\omega(z) = \pm \int_0^{\ell_{\mp,\delta}(z)} \varphi_{\mp t}^* \omega(z) e^{-ts} \, dt \quad \text{for } z \in M_\delta \text{ and } \omega \in \Omega_c^\bullet(M_\delta),$$

which are well defined as operators from $\Omega_c^\bullet(M_\delta)$ to $C(M_\delta, \bigwedge^\bullet T^* M_\delta)$ whenever $\text{Re}(s) \gg 1$, where $C(M_\delta, \bigwedge^\bullet T^* M_\delta)$ denotes the space of continuous differential forms on M_δ. Note that our convention of $R_{\pm,\delta}(s)$ differs from that of [18]. The operator $R_{+,\delta}(s)$ (resp. $R_{-,\delta}(s)$) is the resolvent of $\mathcal{L}_X$ in the future (resp. in the past) for the spectral parameter s. More precisely,

$$(2\text{-}6) \qquad (\mathcal{L}_X \pm s)R_{\pm,\delta}(s) = \text{Id}_{\Omega_c^\bullet(M_\delta)},$$

and for any $(u, v) \in \Omega_c^\bullet(M_\delta \setminus \Gamma_{-,\delta}) \times \Omega_c^\bullet(M_\delta \setminus \Gamma_{+,\delta})$,

$$(2\text{-}7) \qquad \int_{M_\delta} (R_{+,\delta}(s)u) \wedge v = -\int_{M_\delta} u \wedge R_{-,\delta}(s)v.$$

Indeed, for such u and v, there is $L > 0$ such that

$$(2\text{-}8) \qquad \text{supp}(u) \subset \{\ell_{+,\delta} \leqslant L\} \quad \text{and} \quad \text{supp}(v) \subset \{\ell_{-,\delta} \leqslant L\}.$$

In particular, the forms $R_{+,\delta}(s)u$ and $R_{-,\delta}(s)v$ are smooth up to the boundary of M_δ. Indeed, (2-8) implies that, for any $z \in M_\delta$ and $t \in [0, \ell_{-,\delta}(z)]$,

$$\varphi_{-t}^* u(z) \neq 0 \implies t \leqslant L.$$

Therefore, for any $z \in M_\delta$,

$$R_{+,\delta}(s)u(z) = \int_0^{\ell_{-,\delta}(z)} \varphi_{-t}^* u(z) e^{-ts} \, dt = \int_0^{\min(\ell_{-,\delta}(z), L+1)} \varphi_{-t}^* u(z) e^{-ts} \, dt,$$

and thus $R_{+,\delta}u$ is smooth, since $\varphi_{-t}^* u(z) = 0$ if $L \leqslant t \leqslant \ell_{-,\delta}(z)$. Similarly, $R_{-,\delta}(s)v$ is smooth. Finally, note that $\text{supp}(R_{+,\delta}(s)u) \cap \partial M_\delta \subset \partial_+ M_\delta$ and $\text{supp}(R_{-,\delta}(s)v) \cap \partial M_\delta \subset \partial_- M_\delta$. In particular, Stokes' formula and (2-6) imply (2-7).

Because the boundary of Σ_δ is strictly convex, it follows from [15, Proposition 6.1] that the family of operators $R_\pm(s)$ extends to a meromorphic family of operators

$$R_{\pm,\delta}(s): \Omega_c^\bullet(M_\delta) \to \mathcal{D}'^\bullet(M_\delta)$$

satisfying

(2-9)
$$\mathrm{WF}'(R_{\pm,\delta}(s)) \subset \Delta(T^*M_\delta) \cup \Upsilon_{\pm,\delta} \cup (E_{\pm,\delta}^* \times E_{\mp,\delta}^*),$$

where $\Delta(T^*M_\delta)$ is the diagonal in $T^*M_\delta \times T^*M_\delta$,

$$\Upsilon_{\pm,\delta} = \{(\Phi_t(z,\xi),(z,\xi)) \in T^*(M_\delta \times M_\delta) : 0 \leqslant \pm t \leqslant \ell_{\pm,\delta}(z) \text{ and } \langle X(z), \xi \rangle = 0\},$$

and where

$$E_{+,\delta}^* = E_u^*|_{\Gamma_\delta^+}, \quad E_{-,\delta}^* = E_s^*|_{\Gamma_\delta^-}.$$

Here, we write

$$\mathrm{WF}'(R_{\pm,\delta}(s)) = \{(z,\xi,z',\xi') \in T^*(M_\delta \times M_\delta) : (z,\xi,z',-\xi') \in \mathrm{WF}(R_{\pm,\delta}(s))\},$$

where WF is the classical Hörmander wavefront set [23, Section 8]. In fact, by (2-9) we mean that $s \mapsto R_\pm(s)$ is meromorphic as a map $\mathbb{C} \to \mathcal{D}'_{\Gamma'_\pm}(M_\delta \times M_\delta)$ — we identify $R_\pm(s)$ and its Schwartz kernel — where $\boldsymbol{\Gamma}_\pm$ is given by the right-hand side of (2-9), $\boldsymbol{\Gamma}'_\pm = \{(z,\xi,z',-\xi') : (z,\xi,z',-\xi') \in \boldsymbol{\Gamma}_\pm\}$, and where

$$\mathcal{D}'_{\boldsymbol{\Gamma}'_\pm}(M_\delta \times M_\delta) = \{R \in \mathcal{D}'(M_\delta \times M_\delta) : \mathrm{WF}(R) \subset \boldsymbol{\Gamma}'_\pm\}$$

is endowed with its natural topology; see [23, Definition 8.2.2].

Near any $s_0 \in \mathbb{C}$, we have the expansion

$$R_{\pm,\delta}(s) = Y_{\pm,\delta}(s) + \sum_{j=1}^{J(s_0)} \frac{(X \pm s_0)^{j-1} \Pi_{\pm,\delta}(s_0)}{(s-s_0)^j},$$

where $Y_{\pm,\delta}(s)$ is holomorphic near $s = s_0$ and $\Pi_{\pm,\delta}(s_0)$ is a finite-rank projector satisfying

$$\mathrm{WF}'(\Pi_{\pm,\delta}(s_0)) \subset E_{\pm,\delta}^* \times E_{\mp,\delta}^* \quad \text{and} \quad \mathrm{supp}(\Pi_{\pm,\delta}(s_0)) \subset \Gamma_\delta^\pm \times \Gamma_\delta^\mp,$$

where we identified $\Pi_{\pm,\delta}(s_0)$ and its Schwartz kernel.

2.7 Restriction of the resolvent on the geodesic boundary

For any $\varepsilon > 0$, define the open sets

$$A_{\pm,\varepsilon} = \{\ell_{\pm,\delta} > \varepsilon\} \cap \{\ell_{\mp,\delta} > 0\} \subset \mathrm{int}(M_\delta),$$

and notice that, if ε is small, $M_{\delta/2} \subset A_{\pm,\varepsilon}$. Then we have diffeomorphisms $\varphi_{\pm\varepsilon}: A_{\pm,\varepsilon} \to A_{\mp,\varepsilon}$, which induce maps

$$\varphi_{\pm\varepsilon}^*: \mathcal{D}'^\bullet(A_{\mp,\varepsilon}) \to \mathcal{D}'^\bullet(A_{\pm,\varepsilon}).$$

Using a slight abuse of notation, we will still denote by $\varphi_{\pm\varepsilon}^* \colon \mathcal{D}'^\bullet(M_\delta) \to \mathcal{D}'^\bullet(A_{\pm,\varepsilon})$ the composition of $\varphi_{\pm\varepsilon}^*$ with the inclusion $\mathcal{D}'^\bullet(M_\delta) \hookrightarrow \mathcal{D}'^\bullet(A_{\mp,\varepsilon})$, which is given by the restriction. Let

$$\partial = \partial(S\Sigma_\star) = \{(x, v) \in M_\delta : x \in \gamma_\star \sqcup \gamma_\star\}$$

and $\partial_0 = S\gamma_\star \sqcup S\gamma_\star \subset \partial$.

Lemma 2.7 *For any $\varepsilon > 0$ small enough, we have*

$$\mathrm{WF}(\varphi_{\mp\varepsilon}^* R_{\pm,\delta}(s)) \cap N^*(\partial \times \partial) = \varnothing,$$

where

$$N^*(\partial \times \partial) = \{(z', \xi', z, \xi) \in T^*(M_\delta \times M_\delta) : \langle \xi', T_{z'}\partial\rangle = \langle \xi, T_z\partial\rangle = 0\}.$$

Proof We prove the statement for $R_{+,\delta}(s)$. By (2-9) and multiplicativity of wavefront sets (see [23, Theorem 8.2.14]),

$$(2\text{-}10) \qquad \mathrm{WF}'(\varphi_{-\varepsilon}^* R_{+,\delta}(s)) \subset \Delta_\varepsilon \cup \Upsilon_{+,\delta}^\varepsilon \cup (E_{+,\delta}^* \times E_{-,\delta}^*),$$

where

$$\Delta_\varepsilon = \{(\Phi_\varepsilon(z, \xi), (z, \xi)) : (z, \xi) \in T^*M_\delta\}$$

and

$$\Upsilon_{+,\delta}^\varepsilon = \{(\Phi_t(z, \xi), (z, \xi)) : \varepsilon \leqslant t \leqslant \ell_{+,\delta}(z), \langle X(z), \xi\rangle = 0\}.$$

Now assume that there is $\Xi = (z', \xi', z, \xi)$ lying in

$$N^*(\partial \times \partial) \cap (\Delta_\varepsilon \cup \Upsilon_{+,\delta}^\varepsilon \cup (E_{+,\delta}^* \times E_{-,\delta}^*)).$$

If $\Xi \in \Delta_\varepsilon$, then necessarily $z, z' \in \partial_0$, because $\varphi_\varepsilon(\partial \setminus \partial_0) \cap \partial = \varnothing$ whenever $\varepsilon > 0$ is smaller than the injectivity radius of the manifold.[1] We thus have $\xi \in N_z^*\partial = \mathbb{R}\beta(z)$ by Remark 2.4; now $\Phi_\varepsilon(\beta(z))$ does not lie in $\mathbb{R}\beta(\varphi_\varepsilon(z))$ by Lemma 2.2, and therefore $\xi = 0$.

If $\Xi \in \Upsilon_{+,\delta}^\varepsilon$, then there is $T \geqslant \varepsilon$ such that $\Phi_T(z, \xi) = (z', \xi')$ with $\langle \xi, X(z)\rangle = 0$. However, by Remark 2.4, if $(z, \xi) \in N_z^*\partial$ and $\langle \xi, X(z)\rangle = 0$, then $z \in \partial_0$. Thus by what precedes, $\xi = 0$.

Finally, (2-1) and Remark 2.4 imply that $N^*\partial \cap E_{\pm,\delta}^* \subset \{0\}$. Thus we have shown that

$$\mathrm{WF}'(\varphi_{-\varepsilon}^* R_{+,\delta}(s)) \cap N^*(\partial \times \partial) = \varnothing,$$

which is equivalent to the conclusion of the lemma.[2] $\qquad\square$

Remark 2.8 This estimate together with [23, Theorem 8.2.4] imply that the operator $\iota^* \iota_X \varphi_{\mp\varepsilon}^* R_{+,\delta}(s)\iota_*$ is well defined and satisfies

$$\mathrm{WF}(\iota^* \iota_X \varphi_{\mp\varepsilon}^* R_{+,\delta}(s)\iota_*) \subset \mathrm{d}(\iota \times \iota)^\top \mathrm{WF}(\varphi_{\mp\varepsilon}^* R_{+,\delta}(s)),$$

---

[1] Let $x \in \partial\Sigma$. If $(x, v) \in \partial \setminus \partial_0$ satisfies that $(y, w) = \varphi_\varepsilon(x, v) \in \partial$, then the exponential map at x is not injective on the closed ball $B(0, \varepsilon) \subset T_x\Sigma$ of radius ε, since $\pi(\varphi_{\varepsilon'}(x, v')) = y$ for some $v' \in S_x\Sigma$ tangent to $\partial\Sigma$ and some $\varepsilon' \in [0, \varepsilon]$. This follows from the fact that $\partial\Sigma$ is totally geodesic.

[2] Since the set $\{(z, \xi, z', \xi') : (z, \xi, z', -\xi') \in N^*(\partial \times \partial)\}$ coincides with $N^*(\partial \times \partial)$, we may use WF or WF' interchangeably.

where $\iota\colon \partial \hookrightarrow M_\delta$ and $\iota \times \iota\colon \partial \times \partial \hookrightarrow M_\delta \times M_\delta$ are the inclusions. Indeed, the Schwartz kernel of $\iota^* \iota_X \varphi^*_{\mp\varepsilon} R_{+,\delta}(s) \iota_*$ coincides with the pullback by $\iota \times \iota$ of the kernel of $\iota_X \varphi^*_{\mp\varepsilon} R_{+,\delta}(s)$. It also follows from [23, Theorem 8.2.14] that the operator $\iota^* \iota_X \varphi^*_{\mp\varepsilon} R_{+,\delta}(s)$ maps

$$\mathcal{D}'^{\bullet}_{N^*\partial}(M_\delta) \to \mathcal{D}'^{\bullet}(\partial)$$

continuously.

Here the pushforward $\iota_*\colon \Omega^{\bullet}(\partial) \to \mathcal{D}'^{\bullet+1}(M_\delta)$ is defined as follows. If $u \in \Omega^k(\partial)$, we define the current $\iota_* u \in \mathcal{D}'^{k+1}(M_\delta)$ by

$$\langle \iota_* u, v \rangle = \int_\partial u \wedge \iota^* v, \quad v \in \Omega^{n-k-1}(M_\delta).$$

3 The scattering operator

In this section we introduce the dynamical scattering operator $\mathcal{S}_\pm(s)$ associated to our problem. By relating the scattering operator to the resolvent described above, we are able to compute its wavefront set. In consequence, the composition $(\chi \mathcal{S}_\pm(s))^n$ is well defined for $\chi \in C_c^\infty(\partial \setminus \partial_0)$, and we give a formula for its flat trace.

For each $x \in \partial \Sigma_\star$, let $\nu(x)$ be the normal outward pointing vector to the boundary of $\Sigma_\star$, and set

$$\partial_\pm = \{(x,v) \in \partial : \pm\langle \nu(x), v \rangle_g > 0\}.$$

3.1 First definitions

We define the exit times in the future and in the past by

$$\ell_\pm(z) = \inf\{t > 0 : \varphi_{\pm t}(z) \in \partial\} \quad \text{for } z \in M \setminus (\partial_\pm \cup \partial_0),$$

and we declare that $\ell_\pm(z) = \infty$ whenever $z \in \partial_\pm \cup \partial_0$. Then we set

$$\Gamma_\pm = \{z \in M : \ell_\mp(z) = +\infty\}.$$

The set Γ_+ (resp. Γ_-) is the set of points of M which are trapped in the past (resp. in the future). The scattering map $S_\pm\colon \partial_\mp \setminus \Gamma_\mp \to \partial_\pm \setminus \Gamma_\pm$ is defined by

$$S_\pm(z) = \varphi_{\pm\ell_\pm(z)}(z) \quad \text{for } z \in \partial_\mp \setminus \Gamma_\mp,$$

and satisfies $S_\pm \circ S_\mp = \mathrm{Id}_{\partial_\pm \setminus \Gamma_\pm}$. For $s \in \mathbb{C}$, the scattering operator

$$\mathcal{S}_\pm(s)\colon \Omega_c^{\bullet}(\partial_\mp \setminus \Gamma_\mp) \to \Omega_c^{\bullet}(\partial_\pm \setminus \Gamma_\pm)$$

is given by

$$\mathcal{S}_\pm(s)\omega = (S_\mp^*\omega)e^{-s\ell_\mp(\cdot)} \quad \text{for } \omega \in \Omega_c^{\bullet}(\partial_\mp \setminus \Gamma_\mp).$$

Remark 3.1 If $\mathrm{Re}(s)$ is large enough, $\mathcal{S}_\pm(s)$ extends as a map

$$C^0(\partial, \textstyle\bigwedge^{\bullet} T^*\partial) \to C^0(\partial, \textstyle\bigwedge^{\bullet} T^*\partial),$$

where $C^0(\partial, \bigwedge^\bullet T^*\partial)$ is the space of continuous forms on ∂, by declaring that

$$\mathcal{S}_\pm(s)\omega(z) = S_\mp^*\omega(z)e^{-s\ell_\mp(z)} \quad \text{if } z \in \partial_\pm \setminus \Gamma_\pm$$

and $\mathcal{S}_\pm(s)\omega(z) = 0$ otherwise. Indeed, by [Lemma 3.8](#) and [(3-16)](#), there is $C > 0$ such that

$$\|S_\mp^*\omega(z)\| \leqslant Ce^{C\ell_\mp(z)}\|\omega\|_\infty \quad \text{for } z \in \partial_\pm \setminus \Gamma_\pm \text{ and } \omega \in \Omega^\bullet(M),$$

where $\|\omega\|_\infty$ is the uniform norm on $C^0(M, \bigwedge^\bullet T^*M)$.

3.2 The scattering operator via the resolvent

In this section we will see that $\mathcal{S}_\pm(s)$ can be computed in terms of the resolvent. More precisely, we have the following result:

Proposition 3.2 *For any* $\mathrm{Re}(s)$ *large enough,*

$$\mathcal{S}_\pm(s) = (-1)^N e^{\pm\varepsilon s}\iota^*\iota_X\varphi_{\mp\varepsilon}^* R_{\pm,\delta}(s)\iota_*$$

as maps $\Omega_c^\bullet(\partial \setminus \partial_0) \to \mathcal{D}'^\bullet(\partial)$, *where* $N: \Omega^\bullet(\partial) \to \mathbb{N}$ *is the degree operator. That is,* $N(w) = k$ *if* w *is a* k–*form.*

As a consequence of this proposition, [Remark 2.8](#) and the continuity of the pullback [23, Theorem 8.2.4],

$$(\iota \times \iota)^*: \mathcal{D}'^\bullet_{\mathbf{\Gamma}_{\pm,\varepsilon}}(M_\delta \times M_\delta) \to \mathcal{D}'^\bullet(\partial \times \partial),$$

where $\mathbf{\Gamma}_{\pm,\varepsilon}$ is the right-hand side of [(2-10)](#), we get:

Corollary *The scattering operator* $s \mapsto \mathcal{S}_\pm(s): \Omega^\bullet(\partial \setminus \partial_0) \to \mathcal{D}'^\bullet(\partial)$ *extends as a meromorphic family of* $s \in \mathbb{C}$ *with poles of finite rank, with poles contained in the set of Pollicott–Ruelle resonances of* $\mathcal{L}_X$, *that is, the set of poles of* $s \mapsto R_{\pm,\delta}(s)$.

Before proving [Proposition 3.2](#), we start with an intermediate result:

Lemma 3.3 *We have* $\mathcal{S}_\pm(s) = (-1)^N e^{\pm\varepsilon s}\iota^*\iota_X\varphi_{\mp\varepsilon}^* R_{\pm,\delta}(s)\iota_*$ *as maps*

$$\Omega_c^\bullet(\partial_\mp \setminus \Gamma_\mp) \to \mathcal{D}'^\bullet(\partial_\pm \setminus \Gamma_\pm).$$

Remark 3.4 (i) [Proposition 3.2](#) is not a direct consequence of [Lemma 3.3](#). Indeed, the operator $\mathcal{Q}_{\varepsilon,\pm}(s) = (-1)^N e^{\pm\varepsilon s}\iota^*\iota_X\varphi_{\mp\varepsilon}^* R_{\pm,\delta}(s)\iota_*$ could hide some singularities near $\Gamma_\pm$; [Proposition 3.2](#) tells us that this is not the case, at least far from ∂_0.

(ii) A consequence of [Proposition 3.2](#) is that $\mathcal{Q}_{\varepsilon,\pm}(s)$ is identically zero on $\partial_\pm$ (in the sense that $\mathcal{Q}_{\varepsilon,\pm}(s)u = 0$ whenever $\mathrm{supp}(u) \subset \partial_\pm$), as is the case for $\mathcal{S}_\pm(s)$. This can be seen directly from using the fact that

$$\mathrm{supp}(\varphi_{\mp\varepsilon}^* R_{\pm,\delta}(s)\iota_* u) \subset \{\varphi_t(z) : z \in \mathrm{supp}(u) \text{ and } \varepsilon \leqslant \pm t \leqslant \ell_{\pm,\delta}(z)\}.$$

Proof Let $u \in \Omega_c^\bullet(\partial_- \setminus \Gamma_-)$, and $U' \subset \partial_-$ be a neighborhood of supp u such that $\overline{U}'$ does not intersect ∂_0. Let $\varepsilon > 0$ be small enough that

$$z \in \partial_- \implies \ell_+(z) > \varepsilon.$$

The existence of such an ε follows from the fact that, for each $x \in \partial\Sigma$, the exponential map $\exp_x : T_x\Sigma \to \Sigma$ is injective on $B(0, \varepsilon) \subset T_x\Sigma$ whenever $\varepsilon > 0$ is small enough (independent of x). Note also that, for every $z \in \partial_-$,

$$\pi(\varphi_t(z)) \in \Sigma_\delta \setminus \Sigma_\star \quad \text{for} \quad -\ell_{-,\delta}(z) < t < 0,$$

by Lemma 2.6. Next, let us set

$$U = \{(t, z) \in \mathbb{R} \times U' : -\ell_{-,\delta}(z) < t < \varepsilon\}.$$

Then U is diffeomorphic to a tubular neighborhood of U' in M_δ via $(t, z) \mapsto \varphi_t(z)$.[3] Let $\chi \in C^\infty(\mathbb{R})$ be such that $\chi \equiv 1$ near $]-\infty, 0]$ and $\chi \equiv 0$ on $]\frac{1}{2}\varepsilon, +\infty[$. Set, in the above coordinates,

$$\psi(t, z) = \chi(t) e^{-ts} u(z) \in \textstyle\bigwedge^\bullet T^*_{(t,z)} M_\delta,$$

where we see $u(z)$ as a form in $T^*_{(t,z)}M$ by declaring $\iota_{\partial_t} u(z) = 0$. We extend ψ by 0 on M, and we set

$$\phi = \psi - R_{+,\delta}(s)(\mathcal{L}_X + s)\psi.$$

Then ϕ is smooth by (2-5), since supp $\psi \cap \Gamma_- = \varnothing$. Moreover $(\mathcal{L}_X + s)\phi = 0$, and we have

$$\phi|_{\partial_-} = u \quad \text{and} \quad \phi|_{\partial_+} = \mathcal{S}_+(s)u,$$

where $\mathcal{S}_+(s) = \mathcal{S}_+(s)|_{\Omega_c^\bullet(\partial_- \setminus \Gamma_-)}$. Let $h \in \Omega_c^\bullet(M_\delta \setminus \Gamma_{+,\delta})$, so that $R_{-,\delta}(s)h$ is smooth (see the discussion following (2-7)). We have, by (2-6) and (2-7),

$$\int_{M_\delta} \phi \wedge h = \int_{M_\delta} \psi \wedge h - \int_{M_\delta} R_{+,\delta}(s)(\mathcal{L}_X + s)\psi \wedge h = \int_{M_\delta} \psi \wedge h + \int_{M_\delta} (\mathcal{L}_X + s)\psi \wedge R_{-,\delta}(s)h$$

$$= \int_{M_\delta} \psi \wedge h - \int_{M_\delta} \psi \wedge (\mathcal{L}_X - s)R_{-,\delta}(s)h + \int_{\partial M_\delta} \iota_X(\psi \wedge R_{-,\delta}(s)h)$$

$$= \int_{\partial M_\delta} \iota_X(\psi \wedge R_{-,\delta}(s)h) = (-1)^{\deg \psi} \int_{\partial_{-,\delta}} \psi \wedge \iota_X R_{-,\delta}(s)h,$$

since $\iota_X\psi = 0$ and ψ has no support near $\partial_{+,\delta}$. Now we let $\Phi: \partial_- \to \partial_{-,\delta}$ be defined by $\Phi(z) = \varphi_{-\ell_{-,\delta}(z)}(z)$. Assume that the support of h does not intersect U. Then a change of variable gives

$$\Phi^*(\iota_X R_{-,\delta}(s)h)|_{\partial_{-,\delta}} = \iota_X R_{-,\delta}(s)h e^{-s\ell_{-,\delta}(\cdot)}.$$

As we have $\Phi^*(\psi|_{\partial_{-,\delta}}) = (\psi|_{\partial_-})e^{+s\ell_{-,\delta}(\cdot)} = ue^{+s\ell_{-,\delta}(\cdot)}$ by definition of ψ, we obtain

$$(3\text{-}1) \qquad \int_{M_\delta} \phi \wedge h = (-1)^{\deg u} \int_{\partial_-} u \wedge \iota^*(\iota_X R_{-,\delta}(s)h).$$

Now because $(\mathcal{L}_X - s)R_{-,\delta}(s)h = h$, we get $(\mathcal{L}_X - s)R_{-,\delta}(s)h = 0$ near U, and thus $\varphi_\varepsilon^* R_{-,\delta}(s)h = e^{\varepsilon s} R_{-,\delta}(s)h$ near U. Let $v \in \Omega_c^\bullet(\partial_+ \setminus \Gamma_+)$. Then $\overline{U} \cap \operatorname{supp}(v) = \varnothing$ (because $\operatorname{supp}(v) \subset \partial_+ \setminus \Gamma_+$). As

---

[3]The map $G: (t, z) \mapsto \varphi_t(z)$ is clearly smooth on U. By Lemma 2.6, $t \mapsto \rho(\varphi_t(z))$ is strictly increasing for $z \in \partial_-$. Therefore, by uniqueness of the integral curves of X, we see that G is injective. The inverse of G is given by $G^{-1}(z') = (t(z'), z(z'))$, where $t(z') = \inf\{t \geq 0 : \varphi_t(z') \in \partial\}$ and $z(z') = \varphi_{-t(z')}(z')$, which is smooth on $G(U)$ by the implicit function theorem.

$\mathrm{WF}(\iota_* v) \subset N^* \partial$, we may find $h_n \in \Omega_c^\bullet(M_\delta \setminus \Gamma_{+,\delta})$, for $n \in \mathbb{N}$, such that $h_n \to \iota_* v$ in $\mathcal{D}'^\bullet_{N^*\partial}(M_\delta)$, and with the property that $\mathrm{supp}(h_n) \cap \overline{U} = \varnothing.$[4] Then applying (3-1) to $h = h_n$ and letting $n \to \infty$ yields[5]

$$\int_{\partial_+} (\mathcal{S}_+(s)u) \wedge v = (-1)^{\deg u} e^{-\varepsilon s} \int_{\partial_-} u \wedge \iota^* \iota_X \varphi_\varepsilon^* R_{-,\delta}(s) \iota_* v,$$

because $\phi|_{\partial_+} = \mathcal{S}_+(s)u$. Since $\int_{\partial_+} \mathcal{S}_+(s)u \wedge v = \int_{\partial_-} u \wedge \mathcal{S}_-(s)v$, we obtain

$$\mathcal{S}_-(s) = (-1)^{\deg u} e^{-\varepsilon s} \iota^* \iota_X \varphi_\varepsilon^* R_{-,\delta}(s) \iota_*$$

as maps $\Omega_c^\bullet(\partial_+ \setminus \Gamma_+) \to \Omega_c^\bullet(\partial_- \setminus \Gamma_-)$. We can replace X by $-X$ to obtain the desired formula for $\mathcal{S}_+(s)$. □

Proof of Proposition 3.2 Let $u \in \Omega^\bullet(\partial \setminus \partial_0)$ and write $u = u(\tau, \theta) \in T^*_{(\tau, \theta)} \partial$. Let $\chi \in C_c^\infty(\mathbb{R}, [0, 1])$ be such that $\int_{\mathbb{R}} \chi = 1$, $\chi(0) \neq 0$, $\chi \equiv 0$ on $\mathbb{R} \setminus]-\frac{1}{2}\delta, \frac{1}{2}\delta[$ and $\chi > 0$ on $]-\frac{1}{2}\delta, \frac{1}{2}\delta[$. For $n \in \mathbb{N}_{\geqslant 1}$ we set $\chi_n = n\chi(n \cdot)$, so that χ_n converges to the Dirac measure on $\mathbb{R}$ as $n \to +\infty$. We define $u_n \in \Omega_c^\bullet(M_\delta)$ in the (τ, ρ, θ) coordinates by

$$u_n = \chi_n(\rho)u(\tau, \theta) \wedge d\rho.$$

Then $u_n \to (-1)^N \iota_* u$ in $\mathcal{D}'_{N^*\partial}(M_\delta)$, since $\partial = \{\rho = 0\}$. In particular, setting

$$f_n = \iota^* \varphi_{-\varepsilon}^* \iota_X R_{+,\delta}(s)u_n \quad \text{for } n \geqslant 1,$$

Remark 2.8 gives that $f_n \to (-1)^N \iota^* \varphi_{-\varepsilon}^* \iota_X R_{+,\delta}(s) \iota_* u$ in $\mathcal{D}'^\bullet(\partial)$. Moreover, if $\mathrm{Re}(s)$ is large enough, then for any $n \in \mathbb{N}$, we have $(-1)^N \iota^* \varphi_{-\varepsilon}^* \iota_X R_{+,\delta}(s)u_n \in C^0(M_\delta, \bigwedge^\bullet T^* M_\delta)$ and thus $f_n \in C^0(\partial, \bigwedge^\bullet T^* \partial)$. Then we claim that $f_n \to \mathcal{S}_+(s)u$ is in $\mathcal{D}'^\bullet(\partial \setminus \partial_0)$ when $n \to +\infty$, where we recall that

$$\mathcal{S}_+(s)u(z) = \begin{cases} S_-^* u(z)e^{-s\ell_-(z)} & \text{if } z \in \partial_+ \setminus \Gamma_+, \\ 0 & \text{if not.} \end{cases}$$

Let $F = \{|\rho| \leqslant \frac{1}{2}\delta\}$. Since the neighborhood $\{|\rho| < \frac{1}{2}\delta\}$ is strictly convex, there exists $L > 0$ such that, for any $z \in F$ and $T > 0$ with $\varphi_{-T}(z) \in F$, we have

$$(3\text{-}2) \qquad \varphi_{-t}(z) \notin F \quad \text{for all } t \in]0, T[\implies T \geqslant L.$$

Next, take $z \in \partial_+ \setminus \Gamma_+$. Then the set $\{t \in [\varepsilon, \ell_{-,\delta}(z)] : \varphi_{-t}(z) \in F\}$ is a finite union of closed intervals, say

$$\{t \geqslant \varepsilon : \varphi_{-t}(z) \in F\} = \bigcup_{k=0}^{K(z)} [a_k(z), b_k(z)],$$

with $a_k(z) \leqslant b_k(z) \leqslant +\infty$ and $b_k(z) < a_{k+1}(z)$ for every k. We set $\rho(t) = \rho(\varphi_{-t}(z))$ for any $t \geqslant 0$, and we take any smooth norm $\|\cdot\|$ on $\bigwedge^\bullet T^* M_\delta$. Note that $u_n = \chi_n(\rho)u_1$. Moreover, if $z \in M_\delta$ and $t < \ell_{-\delta}(z)$, we have

$$(3\text{-}3) \qquad \|\varphi_{-t}^* u_1(z)\| \leqslant C \|u_1(\varphi_{-t}z)\| \exp(C|t|)$$

---

[4]For example, we may take $h_n(\rho, \tau, \theta) = \chi_n(\rho)v(\tau, \theta) \wedge d\rho$, where $\chi_n \in C_c^\infty(]-\delta, \delta[)$ converges to the Dirac measure.

[5]Here we use that $\iota^* \iota_X \varphi_\varepsilon^* R_{-,\delta}(s)h_n \to \iota^* \iota_X \varphi_\varepsilon^* R_{-,\delta}(s)\iota_* v$ in $\mathcal{D}'^\bullet(\partial)$ as $n \to \infty$ by Remark 2.8, since $h_n \to \iota_* v$ in $\mathcal{D}'^\bullet_{N^*\partial}(M_\delta)$.

for some $C > 0$. Let $\theta_0 > 0$ small and $h \in C^\infty(M_\delta, [0, 1])$ such that $h = 1$ on $\operatorname{supp} u_1$ and

$$(3\text{-}4) \qquad\qquad h(\tau, \rho, \theta) = 0 \quad \text{when} \ \operatorname{dist}(\theta, \pi \mathbb{Z}) < \theta_0.$$

(Such an h exists if θ_0 is small enough, since $u \in \Omega^\bullet(\partial \setminus \partial_0)$.) Then there is $c = c(\theta_0) > 0$ such that $|X\rho| \geqslant c$ on $\operatorname{supp} h$, by Lemma 2.3. In particular, if $\operatorname{Re}(s) > C$, then, by (3-3) and (3-4),

$$
\begin{aligned}
\|f_n(z)\| &\leqslant \int_\varepsilon^{\ell_{-,\delta}(z)} (\chi_n \circ \rho)(\varphi_{-t}(z)) \|\varphi_{-t}^*(\iota_X u_1)(z)\| e^{-ts} \, dt \\
&\leqslant C \|u\|_\infty \sum_{k=0}^{K(z)} e^{(C-s)a_k(z)} \int_{a_k(z)}^{b_k(z)} \chi_n(\rho(t)) h(\varphi_{-t}(z)) \, dt \\
&\leqslant C c^{-1} \|u\|_\infty \sum_{k=0}^{K(z)} e^{(C-s)a_k(z)} \int_{a_k(z)}^{b_k(z)} \chi_n(\rho(t)) |X\rho(\varphi_{-t}(z))| \, dt.
\end{aligned}
$$

Of course, for $t < \ell_{-,\delta}(z)$, we have $X\rho(\varphi_{-t}(z)) = \rho'(t)$. Moreover, by Lemma 2.6, $\pm X^2 \rho > 0$ if $\pm \rho > 0$. Thus we may separate each interval $[a_k(z), b_k(z)]$ into two subintervals on which $|\rho'| > 0$, and change variables to get

$$\int_{a_k(z)}^{b_k(z)} \chi_n(\rho(t)) |\rho'(t)| \, dt \leqslant 2 \int_\mathbb{R} \chi_n(\rho) \, d\rho \leqslant 2.$$

By (3-2), $a_k(z) \geqslant kL$ for any k. Therefore we obtain

$$(3\text{-}5) \qquad\qquad \|f_n(z)\| \leqslant \frac{2\|u\|_\infty}{1 - e^{(C-\operatorname{Re}(s))L}} \quad \text{for } z \in \partial_+ \setminus \Gamma_+ \text{ and } n \geqslant 1.$$

Moreover, if $z \in \partial_-$, we have that $t \mapsto \rho(\varphi_{-t}(z))$ is strictly increasing for any $z \in \partial_-$ by Lemma 2.6. Thus we may reproduce the argument made above to obtain that (3-5) also holds for $z \in \partial_-$. Finally, it is shown in [18, Section 2.4] that $\operatorname{Leb}(\Gamma_+ \cap \partial_+) = 0$.[6] In particular, since each f_n is a continuous, (3-5) holds for any $z \in \overline{(\partial_+ \cup \partial_-) \setminus \Gamma_+} = \partial$.

Next, let $v \in \Omega^\bullet(\partial)$. By Lemma 2.6, the set $\{\varphi_{-t}(z) : t \geqslant \varepsilon\}$ is included in $\{\rho \geqslant \rho(\varphi_{-\varepsilon}(z))\}$ for any $z \in \partial_-$. In particular, as $\operatorname{supp}(u_n) \to \partial$ when $n \to \infty$, we have $f_n(z) \to 0$ for $z \in \partial_-$. By dominated convergence we get, as $n \to \infty$,

$$\int_{\partial_-} f_n \wedge v \to 0.$$

Next, let $\eta > 0$, and $\chi_\pm \in C_c^\infty(\partial_\pm \setminus \Gamma_\pm)$ such that

$$(3\text{-}6) \qquad \chi_- \equiv 1 \quad \text{on } \operatorname{supp}(\chi_+ \circ S_+) \qquad \text{and} \qquad \operatorname{vol}(\operatorname{supp}(1 - \chi_+)) < \eta.$$

Such functions exist, as $\operatorname{Leb}(\Gamma_+ \cap \partial) = 0$. We have

$$\int_{\partial_+} f_n \wedge v = \int_{\partial_+} \chi_+ f_n \wedge v + \int_{\partial_+} (1 - \chi_+) f_n \wedge v.$$

---

[6] Actually, Section 2.4 of [18] says that $\operatorname{Leb}(\Gamma_{+,\delta} \cap \partial_{+,\delta}) = 0$. However, $J_\delta : z \mapsto \varphi_{\ell_{+,\delta}(z)}(z)$ realizes a local diffeomorphism $\partial_+ \to J_\delta(\partial_{+,\delta})$, and we have $J_\delta(\Gamma_+) \subset \Gamma_{+,\delta}$.

Note that $f_n = \tilde{f}_n$ on supp χ_+, where $\tilde{f}_n$ is defined exactly as f_n, replacing u by $\tilde{u} = \chi_- u \in \Omega^\bullet(\partial_- \setminus \Gamma_-)$. By Lemma 3.3, $\mathcal{Q}_{\varepsilon,+}(s)\tilde{u} = \mathcal{S}_+(s)\tilde{u}$, and since $\tilde{f}_n \to \mathcal{Q}_{\varepsilon,+}(s)\tilde{u}$, we have

$$\int_{\partial_+} \chi_+ f_n \wedge v = \int_{\partial_+} \chi_+ \tilde{f}_n \wedge v \to \int_{\partial_+} \chi_+ \mathcal{S}_+(s)\tilde{u} \wedge v = \int_{\partial_+} \chi_+ \mathcal{S}_+(s)u \wedge v,$$

where we used that $\mathcal{S}_+(s)u = \mathcal{S}_+(s)\tilde{u}$ on supp χ_+. On the other hand, as the forms f_n are uniformly bounded by (3-5) and the discussion below, there is $C > 0$ such that, for any $n \geqslant 1$,

$$\left| \int_{\partial_+} (1 - \chi_+)\mathcal{S}_+(s)u \wedge v \right| < C\eta \quad \text{and} \quad \left| \int_{\partial_+} (1 - \chi_+)f_n \wedge v \right| < C\eta,$$

where we used the second part of (3-6). Summarizing the above facts, we obtain that, for $n \geqslant 1$ big enough,

$$\left| \int_\partial f_n \wedge v - \int_\partial \mathcal{S}_+(s)u \wedge v \right| \leqslant 4C\eta.$$

Thus, $f_n \to \mathcal{S}_+(s)u$ in $\mathcal{D}'^\bullet(\partial)$. $\qquad\square$

3.3 Composing the scattering maps

Recall that ∂ has two connected components $\partial^{(1)}$ and $\partial^{(2)}$ that we can identify in a natural way. We denote by $\psi : \partial \to \partial$ the map exchanging those components via this identification (in particular, $\psi(\partial_\pm) = \partial_\mp$), and we set

$$\widetilde{\mathcal{S}}_\pm(s) = \psi^* \circ \mathcal{S}_\pm(s).$$

Also we denote by $\Psi = T^*\partial \to T^*\partial$ the symplectic lift of ψ to $T^*\partial$; that is,

$$\Psi(z, \xi) = (\psi(z), \mathrm{d}\psi_z^{-\top}\xi) \quad \text{for } (z, \xi) \in T^*\partial.$$

Lemma 3.5 *Let $\chi \in C_c^\infty(\partial \setminus \partial_0)$. Then for any $n \geqslant 1$, the composition $(\chi\widetilde{\mathcal{S}}_\pm(s)\chi)^n$, which is well defined from $C^0(\partial, \bigwedge^\bullet T^*\partial)$ to $C^0(\partial, \bigwedge^\bullet T^*\partial)$ for $\mathrm{Re}(s)$ large and holomorphic with respect to s by Remark 3.1, admits a meromorphic continuation as a family of operators $\Omega^\bullet(\partial) \to \mathcal{D}'^\bullet(\partial)$.*

Proof We prove the lemma for $\mathcal{S}_+(s)$. First, assume that $n = 2$. According to [23, Theorem 8.2.14], it suffices to show that $A_1 \cap B_1 = \varnothing$, where for $n \geqslant 1$ we set

$$(3\text{-}7) \qquad \begin{aligned} A_n &= \{(z, \xi) : (z', 0, z, \xi) \in \mathrm{WF}'((\chi\widetilde{\mathcal{S}}_\pm(s))^n) \text{ for some } z' \in \partial\}, \\ B_n &= \{(z, \xi) : (z, \xi, z', 0) \in \mathrm{WF}((\chi\widetilde{\mathcal{S}}_\pm(s))^n) \text{ for some } z' \in \partial\}. \end{aligned}$$

By Proposition 3.2 and Remark 2.8,

$$(3\text{-}8) \qquad \mathrm{WF}'(\chi\mathcal{S}_+(s)\chi)|_{\mathrm{supp}(\chi \times \chi)} \subset \mathrm{d}(\iota \times \iota)^\top(\Delta_\varepsilon \cup \Upsilon_{+,\delta}^\varepsilon \cup (E_{+,\delta}^* \times E_{\mp,\delta}^*)),$$

where Δ_ε and $\Upsilon_{+,\delta}^\varepsilon$ are defined as in the proof of Lemma 2.7. Note that in the coordinates of Lemma 2.3, $\iota(z) = (\tau, 0, \theta) \in \partial$ for any $z = (\tau, \theta) \in \partial$, and thus

$$\mathrm{d}\iota^\top(z, \eta) = \eta_\tau \, \mathrm{d}\tau + \eta_\theta \, \mathrm{d}\theta \quad \text{for } \eta = \eta_\tau \, \mathrm{d}\tau + \eta_\rho \, \mathrm{d}\rho + \eta_\theta \, \mathrm{d}\theta \in T_z^* M.$$

As χ is supported far from ∂_0, we have $(\varphi_\varepsilon(z'), z') \notin \partial \times \partial$ for any $z' \in \operatorname{supp} \chi$ (see for example Lemma 2.6), and, for any $\eta \in T^*_{z'} M_\delta$ such that $\langle X(z'), \eta \rangle = 0$, we have

(3-9) $$\mathrm{d}\iota^\top(z', \eta) = 0 \implies \eta = 0$$

by Lemma 2.3, since $\partial_0 = \{(\tau, 0, \theta) : \theta \in \pi\mathbb{Z}\}$. This implies that A_1 is contained in $E^*_{-,\partial}$, while B_1 is contained in $\Psi(E^*_{+,\partial})$ where $E^*_{+,\partial} = (\mathrm{d}\iota)^\top(E^*_{+,\delta})$. Now we claim that $\Psi(E^*_{+,\partial}) \cap E^*_{-,\partial} \subset \{0\}$ far from ∂_0. By Lemma 2.3 and Section 2.3, for any $z = (\tau, 0, \theta) \in \partial^{(j)} \cap \Gamma_\pm$,

$$E^*_{+,\partial}(z) = \mathbb{R}(\mathrm{d}\iota)^\top_z (r_+(z)\beta(z) - \psi(z)) = \mathbb{R}(-\sin(\theta)r_+(z)\,\mathrm{d}\tau - \mathrm{d}\theta),$$

since $\iota(\tau, \theta) = (\tau, 0, \theta)$. Then $r_+(\psi(z)) \neq r_-(z)$ for all z. Indeed, the contrary would mean that $E_s(z') \cap E_u(z') \neq \{0\}$ for some $z' \in M$ (represented by both z and $\psi(z)$ in M_δ), which is not possible. Now we have $\sin(\theta) \neq 0$ for $z \notin \partial_0$. As a consequence, (3-7) is true, since $\operatorname{supp} \chi \cap \partial_0 = \varnothing$. This concludes the case $n = 2$, and by [23, Theorem 8.2.14] we also have the bound

$$\mathrm{WF}'\big((\chi\widetilde{\mathcal{S}}_+(s)\chi)^2\big) \subset \big(\mathrm{WF}'(\chi\widetilde{\mathcal{S}}_+(s)\chi) \circ \mathrm{WF}'(\chi\widetilde{\mathcal{S}}_+(s)\chi)\big) \cup (B_1 \times \underline{0}) \cup (\underline{0} \times A_1),$$

where $\underline{0}$ denotes the zero section in $T^*\partial$, with $A_1 \subset E^*_{-,\partial}$ and $B_1 \subset \Psi(E^*_{+,\partial})$, and where, for any conical subsets $\Upsilon_1, \Upsilon_2 \subset T^*(M \times M)$, we write

$$\Upsilon_1 \circ \Upsilon_2 = \{(x_1, \xi_1, x_2, \xi_2) : (x_1, \xi_1, y, \eta) \in \Upsilon_1 \text{ and } (y, \eta, x_2, \xi_2) \in \Upsilon_2 \text{ for some } (y, \eta)\}.$$

Note that, if we set

$$E^*_{s,\partial_\pm} = \mathrm{d}\iota^\top(E^*_s|_{\partial_\pm}) \quad \text{and} \quad E^*_{u,\partial_\pm} = \mathrm{d}\iota^\top(E^*_u|_{\partial_\pm}),$$

we have $A_1 \subset E^*_{s,\partial_-}$ and $B_1 \subset \Psi(E^*_{u,\partial_+}) = E^*_{u,\partial_-}$.

We proceed by induction, assuming that, for some $n \geq 2$, the composition $(\chi\widetilde{\mathcal{S}}_\pm(s))^n$ is well defined with the bound

(3-10) $$\mathrm{WF}'\big((\chi\widetilde{\mathcal{S}}_+(s))^n\big) \subset \big(\mathrm{WF}'(\chi\widetilde{\mathcal{S}}_+(s)\chi)^{n-1} \circ \mathrm{WF}'(\chi\widetilde{\mathcal{S}}_+(s)\chi)\big) \cup (B_{n-1} \times \underline{0}) \cup (\underline{0} \times A_1),$$

and that $A_{n-1} \subset E^*_{s,\partial_-}$ and $B_{n-1} \subset E^*_{u,\partial_-}$. This formula implies that the set A_n is included in

$$\big\{(z, \xi) \in T^*\partial : (z', 0, z'', \eta) \in \mathrm{WF}'\big((\chi\widetilde{\mathcal{S}}_+(s)\chi)^{n-1}\big) \text{ and } (z'', \eta, z, \xi) \in \mathrm{WF}'(\chi\widetilde{\mathcal{S}}_+(s)\chi) \text{ for some } z', z'' \in \partial\big\}$$

$$\cup A_1.$$

We have $A_{n-1} \subset E^*_{s,\partial_-}$, and note that $\Psi(E^*_{+,\partial}) \subset E^*_{u,\partial_-}$ and $E^*_{u,\partial_-} \cap E^*_{s,\partial_-} = \{0\}$. Moreover, as mentioned above, $\varphi_\varepsilon(z') \notin \partial$ whenever $z' \in \operatorname{supp}(\chi)$. Thus we obtain, by (3-8),

$$A_n \subset \{(z, \xi) : (z'', \eta, z, \xi) \in \mathrm{d}(\iota \times \iota)^\top(\Upsilon^\varepsilon_{+,\delta}) \text{ for some } \eta \in \Psi(E^*_{s,\partial_-})\} \cup A_1.$$

Now suppose $(z'', \eta, z, \xi) \in \mathrm{d}(\iota \times \iota)^\top(\Upsilon^\varepsilon_{+,\delta})$ with $z'', z \in \operatorname{supp} \chi$. Note that $\Psi(E^*_{s,\partial_-}) = E^*_{s,\partial_+}$ and thus, if $\eta \in \Psi(E^*_{s,\partial_-}) \cap \mathrm{d}\iota(z'')^\top \ker X(z'')$, then $\eta = \mathrm{d}\iota(z'')^\top \tilde{\eta}$ for some $\tilde{\eta} \in E^*_s(z'')$ by (3-9). Since E^*_s is preserved by Φ_{-t}, we obtain $(z, \xi) \in \mathrm{d}\iota^\top(E^*_s)$. In particular, this yields $A_n \subset E^*_{s,\partial_-}$. Reversing the roles

of $(\chi \widetilde{\mathcal{S}}_+(s))^{n-1}$ and $\chi \widetilde{\mathcal{S}}_+(s)$ in (3-10), we get that B_n is included in

$$\{(z, \xi) \in T^* \partial : (z, \xi, z', -\eta) \in \mathrm{WF}(\chi \widetilde{\mathcal{S}}_+(s)\chi) \text{ and } (z', \eta, z'', 0) \in \mathrm{WF}\big((\chi \widetilde{\mathcal{S}}_+(s)\chi)^{n-1}\big) \text{ for some } z', z'' \in \partial\}$$
$$\cup B_1.$$

Proceeding as above, one gets $B_n \subset E^*_{u,\partial_-}$. Finally, $B_n \cap A_1 = \varnothing$, since $E^*_{u,\partial_-} \cap E^*_{s,\partial_-}$ on supp χ by (3-9). As a consequence, the composition $(\chi \widetilde{\mathcal{S}}_+(s)\chi)^{n+1} = (\chi \widetilde{\mathcal{S}}_+(s)\chi)^n \circ (\chi \widetilde{\mathcal{S}}_+(s)\chi)$ is well defined by [23, Theorem 8.2.14], and (3-10) holds with n replaced by $n+1$. $\qquad\square$

Remark 3.6 Using (3-10) inductively, one can actually show that $\mathrm{WF}'\big((\chi \widetilde{\mathcal{S}}_+(s)\chi)^n\big)$ is contained in $\mathrm{d}(\hat{\imath} \times \hat{\imath})^\top \widetilde{\Gamma}_{\varepsilon,+}$, where

$$\widetilde{\Gamma}_{\varepsilon,+} = \{(\hat{\Phi}_t(z, \xi), (z, \xi)) : z, \hat{\varphi}_t(z) \in S\Sigma|_{\gamma_\star} \cap \hat{\imath}(\mathrm{supp}\,\chi), \langle X(z), \xi \rangle = 0, t \geqslant \varepsilon\} \cup (E^*_u \times E^*_s)|_{\mathrm{supp}(\chi \times \chi)}.$$

Here (and only here), in order to avoid confusion, we denote by $\hat{\varphi}$ (resp. $\hat{\Phi}_t$) the complete geodesic flow on $M = S\Sigma$ (resp. the symplectic lift of the geodesic flow on T^*M), and by $\hat{\imath} : \partial \to S\Sigma|_{\gamma_\star} \hookrightarrow M$ the identification of both components of ∂.

3.4 The flat trace of the scattering operator

Let $A : \Omega^\bullet(\partial) \to \mathcal{D}'^\bullet(\partial)$ be an operator such that $\mathrm{WF}'(A) \cap \Delta(T^*\partial) = \varnothing$, where $\Delta(T^*\partial)$ is the diagonal in $T^*(\partial \times \partial)$. Then by [23, Theorem 8.2.4], the pullback $\iota_\Delta^* K_A$ is well defined, where $\iota_\Delta : z \mapsto (z, z)$ is the diagonal inclusion and $K_A \in \mathcal{D}'^3(\partial \times \partial)$ is the Schwartz kernel of A, defined by

$$\int_\partial Au \wedge v = \int_{\partial \times \partial} K_A \wedge \pi_1^* u \wedge \pi_2^* v \quad \text{for } u, v \in \Omega^\bullet(\partial),$$

where $\pi_j : \partial \times \partial \to \partial$ is the projection on the jth factor (for $j = 1, 2$). We then define the (super)flat trace of A by

$$-\mathrm{tr}_s^\flat A = \langle \iota_\Delta^* K_A, 1 \rangle.$$

In fact, one can show that

(3-11)
$$-\mathrm{tr}_s^\flat(A) = \sum_{k=0}^2 (-1)^k \, \mathrm{tr}^\flat(A_k),$$

where $\mathrm{tr}^\flat$ is the transversal trace of Atiyah and Bott [3] and A_k is the operator

$$A_k : C^\infty\big(\partial, {\textstyle\bigwedge}^k T^*\partial\big) \to \mathcal{D}'\big(\partial, {\textstyle\bigwedge}^k T^*\partial\big)$$

induced by A on the space of k–forms (see also [16, Section 2.4] for an introduction to the flat trace).

The purpose of this section is to compute the flat trace of $\mathcal{S}_\pm(s)$. In what follows, for any closed geodesic $\gamma : \mathbb{R}/\ell\mathbb{Z} \to \Sigma$, we will write

$$I_\star(\gamma) = \{z \in S\Sigma|_{\gamma_\star} : z = (\gamma(\tau), \dot{\gamma}(\tau)) \text{ for some } \tau \in \mathbb{R}/\ell\mathbb{Z}\}$$

for the set of incidence vectors of γ along $\gamma_\star$, and

$$I_{\star,\pm}(\gamma) = p_\star^{-1}(I_\star(\gamma)) \cap \partial_\mp,$$

where $p_\star : S\Sigma_\star \to S\Sigma$ is the natural projection.

Proposition 3.7 *Let $\chi \in C_c^\infty(\partial \setminus \partial_0)$. For any $n \geq 1$, the operator $(\chi \widetilde{\mathcal{S}}_\pm(s))^n$ has a well-defined flat trace, and for $\operatorname{Re}(s)$ big enough,*

$$(3\text{-}12) \qquad \operatorname{tr}_s^\flat\left((\chi \widetilde{\mathcal{S}}_\pm(s)\chi)^n\right) = n \sum_{i(\gamma,\gamma_\star)=n} \frac{\ell^\#(\gamma)}{\ell(\gamma)} e^{-s\ell(\gamma)} \left(\prod_{z \in I_{\star,\pm}(\gamma)} \chi^2(z) \right)^{\ell(\gamma)/\ell^\#(\gamma)},$$

where the sum runs over all (not necessarily primitive) closed geodesics γ of (Σ, g) such that $i(\gamma, \gamma_\star) = n$. Here $\ell(\gamma)$ is the length of γ and $\ell^\#(\gamma)$ its primitive length.

This formula should be compared with the formula

$$\operatorname{tr}_s^\flat\left((\chi f^* \chi)^n\right) = \sum_{\gamma \in \operatorname{Per}_n(f)} m^\#(\gamma) \operatorname{sgn}(\det(1 - P_\gamma)) \left(\prod_{z \in \gamma} \chi^2(z) \right)^{n/m^\#(\gamma)},$$

which is valid for any smooth Anosov diffeomorphism $f: Z \to Z$ of a closed manifold Z and $\chi \in C^\infty(Z)$. Here $f^*: C^\infty(Z) \to C^\infty(Z)$ is the pullback operator, $\operatorname{Per}_n(f)$ is the set of n–periodic orbits of f, $m^\#(\gamma)$ is the minimal period of γ and P_γ is the linearized Poincaré map of γ (that is, $P_\gamma = df(z)$ for $z \in \gamma$). Note that the above sum is finite, unlike the sum in (3-12). This is due to the fact that $S_\pm$ is singular at $\Gamma_\pm$, which allows $S_\pm$ to have an infinite number of n–periodic points.

Proof The proof that the intersection

$$(3\text{-}13) \qquad \operatorname{WF}'\left((\chi \widetilde{\mathcal{S}}_\pm(s)\chi)^n\right) \cap \Delta(T^*\partial)$$

is empty follows from the estimate in Remark 3.6, since $E_u^* \cap E_s^* = \{0\}$ and $d\hat\imath(z)^\top : \ker X(\hat\imath(z)) \to T_z^*\partial$ is injective for any $z \in \operatorname{supp}(\chi)$.

For any $n \geq 1$, we define the set $\widetilde{\Gamma}_\pm^n \subset \partial$ by

$$\complement\widetilde{\Gamma}_\pm^n = \{z \in \partial : (\widetilde{S}_\pm)^k(z) \text{ is well defined for } k = 1, \ldots, n\},$$

where $\widetilde{S} = \psi \circ S$. Equivalently,

$$\widetilde{\Gamma}_\pm^1 = \Gamma_\pm \quad \text{and} \quad \widetilde{\Gamma}_\pm^{n+1} = \widetilde{\Gamma}_\pm^n \cap (\widetilde{S}_\mp)^n(\Gamma_\pm \setminus \widetilde{\Gamma}_\mp^n)$$

for $n \geq 1$. Also, we set

$$(3\text{-}14) \qquad \tilde\ell_{\pm,n}(z) = \ell_\pm(z) + \ell_\pm(\widetilde{S}_\pm(z)) + \cdots + \ell_\pm(\widetilde{S}_\pm^{n-1}(z)) \quad \text{for } z \in \complement\widetilde{\Gamma}_\pm^n,$$

where $\ell_\pm(z) = \inf\{t > 0 : \varphi_{\pm t}(z) \in \partial\}$, with the convention that $\tilde\ell_{\pm,n}(z) = +\infty$ if $z \in \widetilde{\Gamma}_\pm^n$. We will need the following:

Lemma 3.8 *Let $n \geq 1$. For any $k \geq 1$, there exists $C_{k,n} > 0$ such that*

$$\|d^k \ell_{\pm,n}(z)\| \leq C_{k,n} \exp(C_{k,n} \ell_{\pm,n}(z)) \quad \text{for } z \in \complement\widetilde{\Gamma}_\pm^n.$$

Proof By induction on n, using (3-14) and the fact that $S_\pm(\complement\widetilde{\Gamma}_\pm^n) = \complement\widetilde{\Gamma}_\pm^{n-1}$, we see that the lemma reduces to proving the estimate

$$(3\text{-}15) \qquad \|d^k \ell_\pm(z)\| \leq C_k \exp(C_k \ell_\pm(z)) \quad \text{for } z \in \complement\widetilde{\Gamma}_\pm^1.$$

In what follows, C_k is a constant depending only on k, which may change at each line. First, notice that $\|\mathrm{d}^k \varphi_t(z)\| \leqslant C_k e^{C_k|t|}$ for any $t \in \mathbb{R}$ and $z \in M_\delta$ such that $\varphi_t(z) \in M_\delta$, for some constant C_k; see for example [8, Proposition A.4.1]. Moreover,

$$\mathrm{d}S_\pm(z) = \mathrm{d}[\varphi_{\ell_\pm(z)}](z) + X(S_\pm(z))\,\mathrm{d}\ell_\pm(z) \quad \text{for } z \in \complement\widetilde{\Gamma}^1_\pm.$$

By induction we obtain that, for any k,

$$(3\text{-}16) \quad \|\mathrm{d}^k S_\pm(z)\| \leqslant C_k \exp(C_k \ell_\pm(z)) + C_k \sum_{j=1}^k \|\mathrm{d}^j \ell_\pm(z)\|^{m_j} \quad \text{with } m_j \in \mathbb{N} \text{ for } j = 1, \ldots, k$$

for any $z \in \complement\widetilde{\Gamma}^1_\pm$. Let (τ, ρ, θ) be the coordinates defined near ∂ given by Lemma 2.3. Then $\rho(S_\pm(z)) = 0$ for $z \in \widetilde{\Gamma}^\pm_1$, and thus

$$(3\text{-}17) \quad (X\rho)(S_\pm(z))\,\mathrm{d}\ell_\pm(z) = -\mathrm{d}\rho(S_\pm(z)) \circ \mathrm{d}[\varphi_{\ell_\pm(z)}](z) \quad \text{for } z \in \complement\widetilde{\Gamma}^1_\pm.$$

Let $z \notin \widetilde{\Gamma}^\pm_1$; Lemma 2.3 gives

$$(3\text{-}18) \quad (X\rho)(S_\pm(z)) = \sin\big(\theta(S_\pm(z))\big).$$

Set $z' = S_\pm(z)$, and write $(\tau(t), \rho(t)) = \pi(\varphi_{\mp t}(z'))$, so that $\rho(0) = 0$. By the proof of Lemma 2.6, $t \mapsto |\rho(t)|$ is strictly increasing (indeed $z \notin \widetilde{\Gamma}^\pm_1$ and thus $\dot{\rho}(0) = \pm X\rho(z') \neq 0$), and whenever $|\rho(t)| \leqslant \frac{1}{2}\delta$,

$$(3\text{-}19) \quad \ddot{\rho}(t) = G(\tau(t), \rho(t))$$

for some smooth function $G \in C^\infty\big((\mathbb{R}/\ell_\star\mathbb{Z})_\tau \times \big[-\frac{1}{2}\delta, \frac{1}{2}\delta\big]_\rho\big)$ satisfying $G(\tau, 0) = 0$ and $\partial_\rho G(\tau, \rho) > 0$. If $D = \sup|\partial_\rho G|$, we have $|G(\tau, \rho)| \leqslant D|\rho|$ and thus $|\ddot{\rho}(t)| \leqslant D|\rho(t)|$, with $\rho(0) = \ddot{\rho}(0) = 0$ and $\dot{\rho}(0) = \pm X\rho(S_\pm(z))$. By comparing the solution of (3-19) with the solutions of $\ddot{y}(t) = Dy(t)$, we obtain

$$|\rho(t)| \leqslant |X\rho(z')|\,\mathrm{sh}(Dt).$$

In particular, $|\rho(t)| < \frac{1}{2}\delta$ whenever $|X\rho(S_\pm(z))|\,\mathrm{sh}(Dt) < \frac{1}{2}\delta$, and thus $\mathrm{sh}(D\ell_\mp(z')) \geqslant \frac{1}{2}\delta|X\rho(z')|$. By (3-18), we conclude that there is $C > 0$ such that

$$(3\text{-}20) \quad \big|\sin\big(\theta(S_\pm(z))\big)\big| \geqslant C \exp(-C\ell_\pm(z)) \quad \text{for } z \in \complement\widetilde{\Gamma}^1_\pm.$$

We therefore obtain, for any $z \in \widetilde{\Gamma}^\pm_1$,

$$\|\mathrm{d}\ell_\pm(z)\| \leqslant C^{-1}\exp(C\ell_\pm(z))\|\mathrm{d}\rho(S_\pm(z))\| \cdot \|\mathrm{d}[\varphi_{\ell_\pm(z)}](z)\| \leqslant C e^{C\ell_\pm(z)}.$$

Now, repeatedly using (3-16), (3-17) and (3-20), we obtain (3-15) by induction on k. $\qquad\square$

Consider $\tilde{\chi} \in C^\infty(\mathbb{R}, [0, 1])$ such that $\tilde{\chi} \equiv 1$ on $]-\infty, 1]$ and $\tilde{\chi} \equiv 0$ on $[2, +\infty[$, and set $\tilde{\chi}_L(z) = \tilde{\chi}(\ell_{\pm,n}(z) - L)$ for $z \in \partial$. Then $\tilde{\chi}_L \in C^\infty_c(\partial \setminus \widetilde{\Gamma}^n_\pm)$, and by (3-11) we see that the Atiyah–Bott trace formula [3, Corollary 5.4] reads in our case

$$(3\text{-}21) \quad \langle \iota^*_\Delta K_{\chi,\pm,n}(s), \tilde{\chi}_L \rangle = \sum_{(\widetilde{S}_\mp)^n(z)=z} e^{-s\ell_{\pm,n}(z)} \tilde{\chi}_L(z) \prod_{k=0}^{n-1} \chi^2((\widetilde{S}_\mp)^k(z)),$$

where $K_{\chi,\pm,n}(s)$ is the Schwartz kernel of $(\chi\mathcal{S}_\pm(s))^n$. Indeed, a simple computation (for example in the spirit of [16, Appendix B][7]) shows that, for any diffeomorphism $f: \partial \to \partial$ with isolated nondegenerate fixed points,

$$\text{(3-22)} \qquad \text{tr}^\flat(F_k) = \sum_{f(z)=z} \frac{\text{tr}\bigwedge^k \mathrm{d}f(z)}{|\det(1-\mathrm{d}f(z))|},$$

where $F_k: \Omega^k(\partial) \to \Omega^k(\partial)$ is defined by $F_k\omega = f^*\omega$ and $\bigwedge^k \mathrm{d}f(z)$ is the map induced by $\mathrm{d}f(z)$ on $\bigwedge^k T_z^*\partial$. Since $\sum_k(-1)^k \text{tr}\big(\bigwedge^k \mathrm{d}f(z)\big) = \det(1-\mathrm{d}f(z))$, it holds that

$$\text{(3-23)} \qquad \text{tr}_s^\flat(F) = \sum_k(-1)^{k+1}\text{tr}^\flat(F_k) = -\sum_{f(z)=z} \text{sgn}\det(1-\mathrm{d}f(z)).$$

Now note that $\tilde{\chi}_L(\chi\widetilde{\mathcal{S}}_\pm(s)\chi)^n$ is by definition the operator given by

$$\text{(3-24)} \qquad \omega \mapsto \tilde{\chi}_L(\cdot)\bigg(\prod_{k=1}^n (\chi\circ(\widetilde{S}_\mp)^k)(\chi\circ(\widetilde{S}_\mp)^{k-1})\bigg)e^{-s\ell_{\pm,n}(\cdot)}(\widetilde{S}_\mp)^{n*}w.$$

Moreover, $\text{sgn}\det(1-\mathrm{d}(\widetilde{S}_\mp)^n(z)) = -1$ for any z such that $(\widetilde{S}_\mp)^n(z) = z$. Indeed, for such a z, $\mathrm{d}(\widetilde{S}_\mp)^n(z)$ is conjugated to the linearized Poincaré map

$$P_z = \mathrm{d}(\varphi_{\ell_{\pm,n}(z)})(z)|_{E^u(z)\oplus E^s(z)},$$

which satisfies $\det(1-P_z) < 0$ as the matrix of P_z in the decomposition $E^u(z) \oplus E^s(z)$ reads $\left(\begin{smallmatrix}\lambda & 0 \\ 0 & \lambda^{-1}\end{smallmatrix}\right)$ for some $\lambda > 1$ (since φ_t preserves the volume form $\alpha \wedge \mathrm{d}\alpha$). Finally, by (3-13), the pairing in the left-hand side of (3-21) is well defined; moreover, the proof of (3-22) can be revisited for the operator (3-24) thanks to the introduction of our cutoff functions $\tilde{\chi}_L$ and χ, yielding (3-21).

As $L \to +\infty$, the right-hand side of (3-21) converges to

$$n\sum_{i(\gamma,\gamma_\star)=n} \frac{\ell^\#(\gamma)}{\ell(\gamma)}e^{-s\ell(\gamma)}\bigg(\prod_{z\in I_{\star,\pm}(\gamma)} \chi^2(z)\bigg)^{\ell(\gamma)/\ell^\#(\gamma)},$$

since for any closed geodesic $\gamma: \mathbb{R}/\mathbb{Z} \to \Sigma$ such that $i(\gamma,\gamma_\star) = n$,

$$\#\{z\in\partial : z = (\gamma(\tau),\gamma'(\tau)) \text{ for some } \tau\} = n\frac{\ell^\#(\gamma)}{\ell(\gamma)}.$$

Note that the sum converges whenever $\text{Re}(s)$ is large enough by Margulis' asymptotic formula, given in the introduction. It remains to see that $\langle i_\Delta^* K_{\chi,\pm,n}(s), 1-\tilde{\chi}_L\rangle \to 0$ as $L \to +\infty$. Note that Lemma 3.8 gives

$$\text{(3-25)} \qquad \|\mathrm{d}^k\tilde{\chi}_L\| \leqslant C_k e^{C_k L}.$$

By Remark 3.1, if $s_0 > 0$ is large enough, one has $\mathcal{S}_\pm(s_0): \Omega^\bullet(\partial) \to C^0\big(\partial, \bigwedge^\bullet T^*\partial\big)$. Also, for any $s \in \mathbb{C}$ with $\text{Re}(s) > 0$,

$$\text{(3-26)} \qquad \mathcal{S}_\pm(s_0+s)w = (\mathcal{S}_\pm(s_0)w)e^{-s\ell_\pm(\cdot)} \quad \text{for } w \in \Omega^\bullet(\partial).$$

---

[7]Actually, in the aforementioned reference, the authors deal with flows, but the diffeomorphism case is even simpler.

Let $N \in \mathbb{N}$ such that $\iota_\Delta^* K_{\chi,\pm,n}(s_0)$ extends as a continuous linear form on $C^N(\partial)$. Then applying Lemma 3.8, we see that if $\mathrm{Re}(s)$ is large enough, the function $\exp(-s\ell_{\pm,n}(\cdot))$ lies in $C^N(\partial)$. Thus, the product $e^{-s\ell_{\pm,n}(\cdot)}\iota_\Delta^* K_{\chi,\pm,n}(s_0)$ is well defined and by (3-25) we have

$$
\begin{aligned}
|\langle e^{-s\ell_{\pm,n}(\cdot)}\iota_\Delta^* K_{\chi,\pm,n}(s_0), (1-\tilde{\chi}_L)\rangle| &= |\langle \iota_\Delta^* K_{\chi,\pm,n}(s_0), (1-\tilde{\chi}_L)e^{-s\ell_{\pm,n}(\cdot)}\rangle| \\
&\leqslant C\,\|(1-\tilde{\chi}_L)e^{-s\ell_{\pm,n}(\cdot)}\|_{C^N(\partial)} \leqslant C_N e^{(C_N-\mathrm{Re}(s))L},
\end{aligned}
$$

since $\ell_{\pm,n} \geqslant L$ on $\mathrm{supp}(1-\tilde{\chi}_L)$. Therefore, to obtain that $\langle i_\Delta^* K_{\chi,\pm,n}(s_0+s), 1-\tilde{\chi}_L\rangle \to 0$ as $L \to +\infty$, it suffices to show that

$$
e^{-s\ell_{\pm,n}(\cdot)}\iota_\Delta^* K_{\chi,\pm,n}(s_0) = \iota_\Delta^* K_{\chi,\pm,n}(s_0+s).
$$

This equality is a consequence of (3-26) and Lemma B.1, since we can take s arbitrarily large. □

Recall from Remark 3.6 that $s \mapsto (\chi\tilde{\mathcal{S}}_\pm(s)\chi)^n$ admits a meromorphic continuation in $\mathcal{D}'^3_{\Gamma'_{\varepsilon,\pm}}(\partial\times\partial)$, where $\Gamma'_{\varepsilon,\pm}$ does not intersect the conormal to the diagonal in $\partial\times\partial$. In particular:

Corollary *The function $s \mapsto \eta_{\pm,\chi,n}(s)$ defined for $\mathrm{Re}(s) \gg 1$ by the right-hand side of (3-12) extends to a meromorphic function on the whole complex plane.*

To prove Theorem 1, we wish to use a standard Tauberian argument near the first pole of $\eta_{\pm,\chi,n}$ to obtain the growth of $N(n, L)$. Indeed, it is known (see Section 5) that $s \mapsto R_{\pm,\delta}(s)$ has a simple pole at $s = h_\star$. However, since $\eta_{\pm,\chi,n}$ is given by the trace of the nth self-composition of the restriction of $R_{\pm,\delta}$ to ∂, it is not clear a priori that $\eta_{\pm,\chi,n}$ will have a singularity at $s = h_\star$. In the next section we obtain some a priori bounds on $N(n, L)$; this will imply that $\eta_{\pm,\chi,n}$ indeed has a pole at $s = h_\star$, of order n.

4 A priori bounds on the growth of geodesics with fixed intersection number with $\gamma_\star$

The purpose of this section is to get a priori bounds on $N(1, L)$ — and $N(2, L)$ in the case where $\gamma_\star$ is separating — using Parry and Pollicott's bound for axiom A flows [35].

Choose some point $x_\star \in \gamma_\star$. Let g be the genus of Σ and $(a_1, b_1, \ldots, a_g, b_g)$ be a basis of generators of Σ, so that the fundamental group of Σ is the finitely presented group given by

$$
(4\text{-}1) \qquad \pi_1(\Sigma) = \langle a_1, b_1, \ldots, a_g, b_g, [a_1, b_1]\cdots[a_g, b_g] = 1\rangle,
$$

where we set $\pi_1(\Sigma) = \pi_1(\Sigma, x_\star)$ for some choice of $x_\star \in \gamma_\star$ (see Figure 2 for the case where $\gamma_\star$ is not separating, and Figure 4 otherwise).
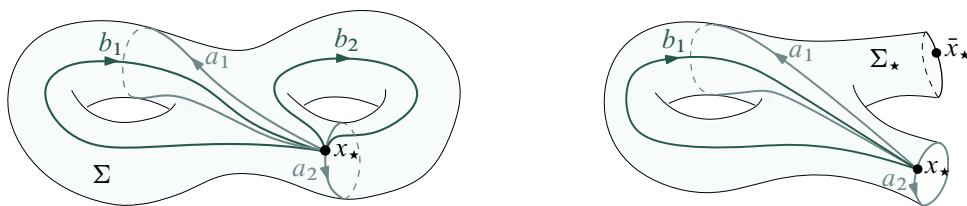
Figure 2: The generators $a_1, b_1, \ldots, a_g, b_g$ of $\pi_1(\Sigma)$ (on the left) and the generators $a_1, b_1, \ldots, a_g$ of $\pi_1(\Sigma_\star)$ (on the right) when $g = 2$. Here $\gamma_\star$ is assumed to be not separating and is represented by a_2 in $\pi_1(\Sigma)$.

4.1 The case $\gamma_\star$ is not separating

Up to applying a diffeomorphism to Σ, we may assume that $\gamma_\star$ is represented by $a_g \in \pi_1(\Sigma)$. The cut surface $\Sigma_\star$ is a topological surface of genus $g - 1$ with 2 punctures, and the fundamental group[8] $\pi_1(\Sigma_\star) = \pi_1(\Sigma_\star, x_\star)$ is the free group given by $\langle a_1, b_1, \ldots, a_g \rangle$, which follows from the fact that $\Sigma_\star$ is homotopically equivalent to a connected sum of $2g - 1$ circles. We refer to Figure 2 for a picture of the generators and the choice of $x_\star$. By the presentation of $\pi_1(\Sigma)$ given above, we have

$$(4\text{-}2) \qquad b_g a_g b_g^{-1} = a_g' \quad \text{where } a_g' = [a_1, b_1] \cdots [a_{g-1}, b_{g-1}] a_g,$$

and note that a_g' also defines an element of $\pi_1(\Sigma_\star)$.

Lemma 4.1 *The map $q_\star \colon \Sigma_\star \to \Sigma$ given by the identification of the boundary components of $\Sigma_\star$ induces a map $q_{\star,*} \colon \pi_1(\Sigma_\star) \to \pi_1(\Sigma)$, which is injective.*

Proof Let $\langle a_g \rangle$ (resp. $\langle a_g' \rangle$) be the infinite cyclic subgroup of $\pi_1(\Sigma_\star)$ generated by a_g (resp. a_g'). Then by (4-1) and (4-2), the group $\pi_1(\Sigma)$ is the HNN[9] extension $\pi_1(\Sigma_\star)*_\phi$ of $\pi_1(\Sigma_\star)$ with respect to the isomorphism $\phi \colon \langle a_g' \rangle \to \langle a_g \rangle$ given by $\phi(a_g') = a_g$, that is, $\pi_1(\Sigma_\star)*_\phi$ is the finitely presented group defined by

$$\pi_1(\Sigma_\star)*_\phi = \langle a_1, b_1, \ldots, a_g, t : t^{-1} a_g' t = a_g \rangle;$$

see [30, Section IV.2]. Now the map $q_{\star,*} \colon \pi_1(\Sigma_\star) \to \pi_1(\Sigma)$ coincides with the natural map $\pi_1(\Sigma_\star) \to \pi_1(\Sigma_\star)*_\phi$, and this map is injective by [30, Theorem IV.2.1]. $\qquad \square$

We may see the cut surface $\Sigma_\star$ as the convex core of a complete, noncompact, negatively curved surface, with funnels. Indeed, by Lemma 4.1, the group $\pi_1(\Sigma_\star)$ can be thought of as a subgroup of $\pi_1(\Sigma)$, and the convex core of the infinite surface $\Sigma_\star^e = \pi_1(\Sigma_\star) \backslash \tilde{\Sigma}$ is canonically isometric to $\Sigma_\star$ (here $\tilde{\Sigma}$ is a universal cover of Σ). Another way to obtain this is by gluing two arbitrary funnels as follows. Recall that near each connected component of the boundary $\partial \Sigma_\star \subset \Sigma_\delta$ we have coordinates

---

[8]Here, in order not to burden the notation, we still denote by $x_\star \in \Sigma_\star$ a lift of $x_\star \in \Sigma$ by the natural map $q_\star \colon \Sigma_\star \to \Sigma$; see Figure 2.

[9]HNN refers to the authors Graham Higman, Bernhard Neumann and Hanna Neumann [22].

$(\tau, \rho) \in \mathbb{R}/\ell_\star \mathbb{Z}_\tau \times [-\delta, \delta]_\rho$ given by Lemma 2.3, for which $\partial \Sigma_\star = \{\rho = 0\}$ and $\partial \Sigma_\delta = \{\rho = \delta\}$. In those coordinates, the metric has the form $d\rho^2 + f(\tau, \rho) \, d\tau^2$ for some smooth function f satisfying $\partial_\rho f(\tau, 0) = 0$ and $\kappa(\tau, \rho) = -\partial_\rho^2 f(\tau, \rho) / f(\tau, \rho)$. Then we arbitrarily extend f to a smooth function on $(\mathbb{R}/\ell_\star \mathbb{Z})_\tau \times [-\delta, +\infty[$ so that, for some constants $c, C > 0$,

$$c \leqslant \frac{\partial_\rho^2 f}{f} \leqslant C.$$

By gluing the funnels $(\mathbb{R}/\ell_\star \mathbb{Z}) \times [0, \infty[$ and $\Sigma_\star$ along the corresponding connected components, we obtain a complete negatively curved surface $\Sigma_\star^e$, whose metric in the funnels is given by $d\rho^2 + f(\tau, \rho) \, d\tau^2$. We will again denote by (φ_t) the geodesic flow on the unit tangent bundle $S\Sigma_\star^e$ of $\Sigma_\star^e$.

Let $\widetilde{\Sigma}_\star$ denote the universal cover of $\Sigma_\star^e$ and let $\tilde{x}_\star \in \widetilde{\Sigma}_\star$ be such that $\pi(\tilde{x}_\star) = x_\star$, where $\pi \colon \widetilde{\Sigma}_\star \to \Sigma_\star^e$ is the natural projection. Then $\pi_1(\Sigma_\star^e, x_\star) = \pi_1(\Sigma_\star)$ acts on $\widetilde{\Sigma}_\star$ by deck transformations so that $\Sigma_\star^e \simeq \pi_1(\Sigma_\star) \backslash \widetilde{\Sigma}_\star$. Moreover, Lemma 2.6 implies that the recurrent set of the geodesic flow on $S\Sigma_\star^e$ is compact and included in $S\Sigma_\star$; thus $\pi_1(\Sigma_\star)$ is convex–cocompact in the sense of [12]. The aforementioned lemma also implies that every closed geodesic in $\Sigma_\star^e$ which is not contained in $\partial \Sigma_\star$ is actually contained in the interior of $\Sigma_\star$.

It is well known that there is a one-to-one correspondence between oriented closed geodesics on $\Sigma_\star^e$ (all of them belonging to $\Sigma_\star$) and the set of free homotopy classes of loops in $\Sigma_\star^e$. The latter set is itself in one-to-one correspondence with the set of conjugacy classes of $\pi_1(\Sigma_\star)$. We set

$$\ell_\star(w) = \operatorname{dist}(\tilde{x}_\star, w\tilde{x}_\star) \quad \text{for } w \in \pi_1(\Sigma_\star),$$

where the distance comes from the metric $\pi^* g$ on $\widetilde{\Sigma}_\star$. For any $w \in \pi_1(\Sigma_\star)$, we denote by $[w]$ the associated conjugacy class of $\pi_1(\Sigma_\star)$. Note that if $\gamma_{[w]}$ denotes the unique geodesic in the free homotopy class of w (which is represented by the conjugacy class $[w]$), we have $\ell(\gamma_{[w]}) \leqslant \ell_\star(w)$. We also denote by

$$(4\text{-}3) \qquad \operatorname{wl}(w) = \min\{n \geqslant 0 : w = \alpha_1 \cdots \alpha_n \text{ with } \alpha_j \in \mathscr{L}_g \setminus \{b_g, b_g^{-1}\}\}$$

the word length of an element $w \in \pi_1(\Sigma_\star)$, where $\mathscr{L}_g = \bigcup_{k=1}^g \{a_k, a_k^{-1}, b_k, b_k^{-1}\}$. We will say that a word $\alpha_1 \cdots \alpha_k$ with $\alpha_j \in \mathscr{L}_g$ is reduced if $\alpha_j \neq (\alpha_{j+1})^{-1}$ for any $j = 1, \ldots, k - 1$. As $\pi_1(\Sigma_\star)$ is free, for each $w \in \pi_1(\Sigma_\star)$, there is exactly one reduced word $\alpha_1 \cdots \alpha_n$ such that $n = \operatorname{wl}(w)$; see [30, page 4]. It follows from the Milnor–Švarc lemma [11, Proposition I.8.19] that, for some constant $D > 0$,

$$(4\text{-}4) \qquad \frac{1}{D} \operatorname{wl}(w) - D \leqslant \ell_\star(w) \leqslant D \operatorname{wl}(w) + D \quad \text{for } w \in \pi_1(\Sigma_\star).$$

Also, as $\pi_1(\Sigma_\star)$ is convex cocompact, we have the classical orbital counting (see [42, paragraphe 1.F and corollaire 2])

$$(4\text{-}5) \qquad \#\{w \in \pi_1(\Sigma_\star) : \ell_\star(w) \leqslant L\} \sim A e^{h_\star L} \quad \text{as } L \to \infty$$

for some $A > 0$, where $h_\star > 0$ is the topological entropy of the geodesic flow of $(\Sigma_\star^e, g)$ restricted to the trapped set

$$K_\star^e = \{(x, v) \in S\Sigma_\star^e : \varphi_t(x, v) \in S\Sigma_\star \text{ for } t \in \mathbb{R}\}.$$

In fact, $h_\star > 0$ also coincides with the entropy of the geodesic flow of (Σ, g) restricted to the trapped set $K_\star$ mentioned in the introduction,

$$K_\star = \overline{\{(x, v) \in S\Sigma : \pi(\varphi_t(x, v)) \in \Sigma \setminus \gamma_\star \text{ for } t \in \mathbb{R}\}},$$

where the closure is taken in $S\Sigma$ and $K_\star^e = p_\star^{-1}(K_\star)$, where $p_\star \colon S\Sigma_\star \to S\Sigma$ is the natural map given by the identification of both components of $\partial S\Sigma_\star$.

4.1.1 Lower bound In this section we will prove:

Proposition 4.2 *If $\gamma_\star$ is not separating, then there is $C > 0$ such that, for any L large enough,*

$$N(1, L) \geqslant C\frac{e^{h_\star L}}{L}.$$

Note that Theorem 1 actually gives $N(1, L) \sim c_\star e^{h_\star L}$, so Proposition 4.2 is not sharp. We could obtain a better bound with the methods presented in Section 4.2, which deals with the separating case; however, Proposition 4.2 will be sufficient for our purposes (see Remarks 5.2, 5.3 and 5.4).

Lemma 4.3 *Take $w, w' \in \pi_1(\Sigma_\star)$. Then $[wb_g] = [w'b_g]$ as conjugacy classes of $\pi_1(\Sigma)$ if and only if $w = a_g^n w' a_g'^{-n}$ in $\pi_1(\Sigma_\star)$ for some $n \in \mathbb{Z}$.*

Proof If $w = a_g^n w' b_g a_g^{-n} b_g^{-1}$, then clearly wb_g and $w'b_g$ are conjugate in $\pi_1(\Sigma, x_\star)$. Reciprocally, assume that $[wb_g] = [w'b_g]$. We may find smooth paths γ and γ' representing respectively the elements wb_g and $w'b_g$, with $i(\gamma, \gamma_\star) = i(\gamma', \gamma_\star) = 1$ and such that the intersections $\gamma \cap \gamma_\star$ and $\gamma' \cap \gamma_\star$ are transverse. As $[wb_g] = [w'b_g]$, the loops γ and γ' lie in the same free homotopy class. Thus there is a smooth homotopy $H \colon [0, 1] \times \mathbb{R}/\mathbb{Z} \to \Sigma$ such that $H(0, \cdot) = \gamma$ and $H(1, \cdot) = \gamma'$. We may assume that H is transverse to $\gamma_\star$ (see for example [20, Corollary, page 73]) in the sense that

$$\mathrm{d}H(s, \tau)(T_{(s,\tau)}([0, 1] \times \mathbb{R}/\mathbb{Z})) + T_{H(s,\tau)}\gamma_\star = T_{H(s,\tau)}\Sigma \quad \text{for } H(s, \tau) \in \gamma_\star.$$

In particular, $H^{-1}(\gamma_\star)$ is a smooth submanifold of $[0, 1] \times \mathbb{R}/\mathbb{Z}$. As γ and γ' intersect $\gamma_\star$ transversally exactly once, $H^{-1}(\gamma_\star) \cap (\{j\} \times \mathbb{R}/\mathbb{Z}) = \{j\} \times \{[0]\}$ for $j = 0, 1$ (here $[0]$ is sent to $x_\star$ by both γ and γ'). Thus, necessarily, there exists an embedding $F \colon [0, 1] \to [0, 1] \times \mathbb{R}/\mathbb{Z}$ such that $\mathrm{Im}(F) \subset H^{-1}(\gamma_\star)$ and $F(j) = (j, [0])$ for $j = 0, 1$ (see Figure 3). Write $F = (S, T)$, and define

$$\widetilde{H}(s, t) = H\big(S(s), [T(s) + t]\big) \quad \text{for } (s, t) \in [0, 1] \times [0, 1].$$

It is immediate to check that $\widetilde{H}$ realizes a homotopy between γ and γ', and we have $\widetilde{H}(s, 0) = H(F(s)) \in \gamma_\star$ for any $s \in [0, 1]$. For any s, let us denote by c_s the path $[0, 1] \ni u \mapsto \widetilde{H}(su, 0)$ which links $x_\star$ to $H\big(S(s), [T(s)]\big)$ within $\gamma_\star$. The continuous family of paths $s \mapsto \gamma_s$, where γ_s is given by the concatenation $c_s^{-1}\widetilde{H}(s, \cdot)c_s$, realizes a continuous interpolation between $\gamma_0 = \gamma$ and $\gamma_1 = c_1^{-1}\gamma'c_1$. As $S(1) = 1$ and $T(1) = [0]$ we have $c_1(0) = c_1(1) = x_\star$, and since $c_1(u) \in \gamma_\star$ for each $u \in [0, 1]$ we get $c_1 = a_g^{-n}$ for some $n \in \mathbb{Z}$. This yields $wb_g = a_g^n w' b_g a_g^{-n}$ in $\pi_1(\Sigma)$, and thus $w = a_g^n w' a_g'^{-n}$, where the equality stands in $\pi_1(\Sigma)$. By Lemma 4.1, this equality actually holds in $\pi_1(\Sigma_\star)$. $\square$
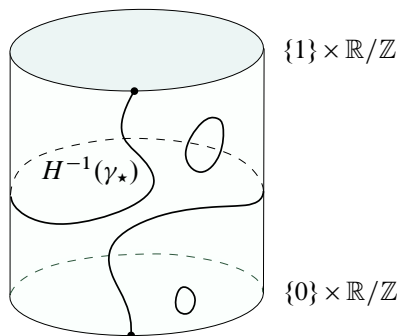
Figure 3: Proof of Lemma 4.3. The path linking $(0, [0]) \in \{0\} \times \mathbb{R}/\mathbb{Z}$ to $(1, [0])$ is the image of F.

Proof of Proposition 4.2 In what follows, C is a constant that may change at each line. For any $w \in \pi_1(\Sigma_\star)$ and $n \in \mathbb{Z}$, by (4-4),

$$(4\text{-}6) \qquad \ell_\star(a_g^n w a_g'^{-n}) \geq \frac{1}{D} \, \text{wl}(a_g^n w a_g'^{-n}) - D.$$

Let w' be the unique reduced word such that $w' = w a_g'^{-n}$. Then write $w' = a_g^{-k} w''$ for some w'', where $|k|$ is maximal, and note that necessarily $|k| \leq \text{wl}(w) + 1$, since $a_g' = [a_1, b_1] \cdots [a_{g-1}, b_{g-1}] a_g$. Then

$$\text{wl}(a_g^n w a_g'^{-n}) = |n| - |k| + \text{wl}(w'') = |n| - 2|k| + \text{wl}(w') \geq |n| - 2(\text{wl}(w) + 1) + \text{wl}(w').$$

Now the triangle inequality for wl gives $(4(g-1) + 1)|n| = \text{wl}(a_g'^{-n}) \leq \text{wl}(w') + \text{wl}(w^{-1})$, and thus we obtain $\text{wl}(a_g^n w a_g'^{-n}) \geq C|n| - C \, \text{wl}(w) - C$ for each n. Injecting this in (4-6) yields (for some different C)

$$\ell_\star(a_g^n w a_g'^{-n}) \geq C|n| - C \, \text{wl}(w) - C \quad \text{for } n \in \mathbb{Z}.$$

In particular, for any L and w such that $\ell_\star(w) \leq L$, by (4-4),

$$(4\text{-}7) \qquad |\{n \in \mathbb{Z} : \ell_\star(a_g^n w a_g'^{-n}) \leq L\}| \leq CL + C.$$

Now, for $w \in \pi_1(\Sigma_\star)$ set $\mathcal{C}_w = \{a_g^n w a_g'^{-n} : n \in \mathbb{Z}\} \subset \pi_1(\Sigma_\star)$, and denote by $\mathscr{C}$ the set $\{\mathcal{C}_w : w \in \pi_1(\Sigma_\star)\}$. For $\mathcal{C} \in \mathscr{C}$, we set $\ell_\star(\mathcal{C}) = \inf_{w \in \mathcal{C}} \ell_\star(w)$. Then by Lemma 4.3, we have a well-defined and injective map

$$\{\mathcal{C} \in \mathscr{C} : \ell_\star(\mathcal{C}) \leq L\} \to \{\gamma \in \mathcal{P}_1 : \ell(\gamma) \leq L + C\}, \quad \mathcal{C}_w \mapsto [w b_g],$$

where $\mathcal{P}_1$ denotes the set of primitive geodesics γ such that $i(\gamma, \gamma_\star) = 1$.[10] In particular we get, with (4-7) and (4-5),

$$(4\text{-}8) \qquad N(1, L) \geq |\{\mathcal{C} \in \mathscr{C} : \ell_\star(\mathcal{C}) \leq L - C\}| \geq \frac{1}{CL + C} \sum_{\substack{\mathcal{C} \in \mathscr{C} \\ \ell_\star(\mathcal{C}) \leq L - C}} |\{w \in \mathcal{C} : \ell_\star(w) \leq L - C\}|$$

$$= \frac{1}{CL + C} |\{w \in \pi_1(\Sigma_\star) : \ell_\star(w) \leq L - C\}| \geq \frac{1}{CL + C} \exp(h_\star(L - C)),$$

where the equality comes from the fact that $\pi_1(\Sigma_\star)$ is the disjoint union of the subsets $\mathcal{C}$ with $\mathcal{C} \in \mathscr{C}$. $\square$

---

[10]Each class $[w b_g]$ defines a geodesic in $\mathcal{P}_1$. Indeed, it follows from Lemma 2.1 that $i([w b_g], \gamma_\star) \leq 1$. On the other hand, the absolute value of the algebraic intersection number between $w b_g$ and a_g is 1, and this implies that there is at least one intersection point between $[w b_g]$ and $\gamma_\star$, since the algebraic intersection number is preserved by free homotopies.
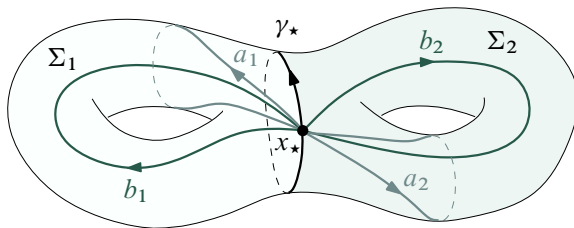
Figure 4: The generators $a_1, b_1, \ldots, a_g, b_g$ of $\pi_1(\Sigma)$. Here $\gamma_\star$ is assumed to be separating and $g_1 = g_2 = 1$.

4.1.2 Upper bound Each $\gamma \in \mathcal{P}_1$ with $\ell(\gamma) \leq L$ lies in the free homotopy class of $w' b_g^{\pm 1}$ for some $w' \in \pi_1(\Sigma_\star, x_\star)$ and $\ell_\star(w) \leq L + C$. In particular, (4-5) gives the bound

$$N(1, L) \leq C \exp(h_\star L)$$

for large L. Now let $\gamma \in \mathcal{P}_2$ with $\ell(\gamma) \leq L$. Then we may find a deformation of the loop γ into a loop γ' which is represented by the conjugacy class of $w b_g^{\pm 1} w' b_g^{\pm 1}$ in $\pi_1(\Sigma)$ for some $w, w' \in \pi_1(\Sigma_\star)$. This deformation can be made so that $\ell_\star(w) + \ell_\star(w') \leq L + C$. Thus,

$$N(2, L) \leq C \sum_{\substack{w, w' \in \pi_1(\Sigma_\star) \\ \ell_\star(w) + \ell_\star(w') \leq L + C}} 1 \leq \sum_{k=0}^{L+C} C \exp(h_\star k) C \exp(h_\star(L + C - k)) \leq C' L \exp(h_\star L).$$

Iterating this process, we finally get, for large L,

$$N(n, L) \leq C L^{n-1} \exp(h_\star L).$$

4.2 The case $\gamma_\star$ is separating

In this section we assume $\gamma_\star$ is separating, and we write $\Sigma \setminus \gamma_\star = \Sigma_1 \sqcup \Sigma_2$, where the surfaces Σ_j are connected. Up to applying a diffeomorphism to Σ, we may assume that $\gamma_\star$ represents the class

(4-9) $$[a_1, b_1] \cdots [a_{g_1}, b_{g_1}] = [a_g, b_g]^{-1} \cdots [a_{g_1+1}, b_{g_1+1}]^{-1} \in \pi_1(\Sigma)$$

(see Figure 4). Here g_1 is the genus of the surface Σ_1, and the genus g_2 of Σ_2 satisfies $g_1 + g_2 = g$.

We set $\pi_1(\Sigma) = \pi_1(\Sigma, x_\star)$ and $\pi_1(\Sigma_j) = \pi_1(\Sigma_j, x_\star)$ for $j = 1, 2$ (we see Σ_j as a compact surface with boundary $\gamma_\star$ so that $x_\star$ lives on both surfaces). Then $\pi_1(\Sigma_1)$ and $\pi_1(\Sigma_2)$ are the free groups generated by $a_1, b_1, \ldots, a_{g_1}, b_{g_1}$ and $a_{g_1+1}, b_{g_1+1}, \ldots, a_g, b_g$, respectively, and we denote by $w_{\star,1}$ and $w_{\star,2}$ the two natural words given by (4-9) representing $\gamma_\star$ in $\pi_1(\Sigma_1)$ and $\pi_1(\Sigma_2)$, respectively. Note that we have a well-defined map

$$\pi_1(\Sigma_1) \times \pi_1(\Sigma_2) \to \pi_1(\Sigma), \quad (w_1, w_2) \mapsto w_2 w_1,$$

given by the composition of two curves.

Lemma 4.4 *For $j = 1, 2$, the map $q_{j,*} \colon \pi_1(\Sigma_j) \to \pi_1(\Sigma)$ induced by the inclusion $\Sigma_j \hookrightarrow \Sigma$ is injective.*

Proof For $j = 1, 2$ let $\langle w_{\star,j} \rangle$ be the infinite cyclic group of $\pi_1(\Sigma_j)$ generated by $w_{\star,j}$, and let $\phi \colon \langle w_{\star,1} \rangle \to \langle w_{\star,2} \rangle$ be the isomorphism given by $\phi(w_{\star,1}) = w_{\star,2}$. By (4-1), the group $\pi_1(\Sigma)$ is the free product with amalgamation $\pi_1(\Sigma_1) *_\phi \pi_1(\Sigma_2)$, that is, the finitely presented group given by

$$\pi_1(\Sigma_1) *_\phi \pi_1(\Sigma_2) = \{a_1, b_1, \ldots, a_g, b_g : w_{\star,1} = \phi(w_{\star,1})\};$$

see [30, Section IV.2]. With this representation, the map $q_{j,*}$ coincides with the natural map $\pi_1(\Sigma_j) \to \pi_1(\Sigma_1) *_\phi \pi_1(\Sigma_2)$, which is injective by [30, Theorem IV.2.6]. $\qquad\square$

For any $w \in \pi_1(\Sigma)$, we will denote by $[w]$ its conjugacy class and by γ_w the unique geodesic of Σ such that γ_w is isotopic to any curve in w (in fact we will often identify $[w]$ and γ_w). Let $(\tilde{\Sigma}, \tilde{g})$ be the universal cover of (Σ, g), and choose $\tilde{x}_\star \in \tilde{\Sigma}$ some lift of $x_\star$. Then $\pi_1(\Sigma)$ acts as deck transformations on $\tilde{\Sigma}$ and we will write

$$\ell_\star(w) = \mathrm{dist}_{\tilde{\Sigma}}(\tilde{x}_\star, w\tilde{x}_\star) \quad \text{for } w \in \pi_1(\Sigma).$$

As in the preceding subsection, we have the orbital counting

$$(4\text{-}10) \qquad \#\{w_j \in \pi_1(\Sigma_j) : \ell_\star(w_j) \leqslant L\} \sim A_j e^{h_j L} \quad \text{as } L \to \infty \text{ for } j = 1, 2$$

for some $A_1, A_2 > 0$, where $h_j > 0$ is the topological entropy of the geodesic flow restricted to the trapped set

$$K_j = \overline{\{(x, v) \in S\Sigma_j^\circ : \varphi_t(x, v) \in S\Sigma_j^\circ \text{ for } t \in \mathbb{R}\}},$$

where $\Sigma_j^\circ = \Sigma_j \setminus \partial\Sigma_j$ for $j = 1, 2$.

4.2.1 Lower bound Unlike the case where $\gamma_\star$ is not separating, we will need a better lower bound. Namely, we prove here the following result:

Proposition 4.5 *Assume that $\gamma_\star$ is separating and that $h_1 = h_2 = h_\star$. Then there is $C > 0$ such that, for L large enough,*

$$(4\text{-}11) \qquad\qquad\qquad N(2, L) \geqslant \frac{CLe^{h_\star L}}{\log(L)^4}.$$

If $h_1 \neq h_2$ we have, for L large enough and $h_\star = \max(h_1, h_2)$,

$$(4\text{-}12) \qquad\qquad\qquad N(2, L) \geqslant \frac{Ce^{h_\star L}}{\log(L)^2}.$$

Note that Theorem 2 gives $N(2, L) \sim CLe^{h_\star L}$ if $h_1 = h_2$ and $N(2, L) \sim Ce^{h_\star L}$ if $h_1 \neq h_2$. In particular, Proposition 4.5 gives a bound which is sharp up to a logarithmic loss, whereas in Proposition 4.2, we had a linear loss. Indeed, obtaining a sharper bound is important here, because a linear defect would not be sufficient to obtain Theorem 2 in the case $h_1 = h_2$ — at least with our methods. If $h_1 \neq h_2$, a linear loss would nevertheless be sufficient, but our proof of (4-11) actually gives (4-12) without too much effort. We refer to Remarks 5.2, 5.3 and 5.4 for a more detailed discussion about the importance of (4-11).

The strategy to prove Proposition 4.5 is the following. We wish to construct enough closed geodesics intersecting $\gamma_\star$ exactly twice by considering conjugacy classes of the form $[w_2 w_1]$ where $w_j \in \pi_1(\Sigma_j)$ for $j = 1, 2$. Lemma 4.6 will tell us that, if w_j is not a power of $w_{\star,j}$ for $j = 1, 2$, then the closed geodesic representing $[w_2 w_1]$ indeed intersects $\gamma_\star$ exactly twice. Next, in Lemma 4.7, we describe the injectivity defect of the map $(w_1, w_2) \mapsto [w_2 w_1]$. Finally, in Proposition 4.8, we show that this injectivity defect is not too harmful in the sense that there are not too many $w_j, w'_j \in \pi_1(\Sigma_j)$ such that $[w_2 w_1] = [w'_2 w'_1]$. This will allow us to obtain the desired bound with a logarithmic loss.

Lemma 4.6 *For two elements $w_j \in \pi_1(\Sigma_j)$ for $j = 1, 2$, we have $i(\gamma_{w_2 w_1}, \gamma_\star) = 2$ except if $w_j = w^k_{\star,j}$ in $\pi_1(\Sigma_j)$ for some $k \in \mathbb{Z}$ and $j \in \{1, 2\}$, in which case $i(\gamma_{w_2 w_1}, \gamma_\star) = 0$.*

Proof Let $\gamma \colon \mathbb{R}/\mathbb{Z} \to \Sigma$ be a smooth curve in the free homotopy class of $w_2 w_1$ such that

$$\{\tau \in \mathbb{R}/\mathbb{Z} : \gamma(\tau) \in \gamma_\star\} = \{\tau_1, \tau_2\} \quad \text{for some } \tau_1 \neq \tau_2 \in \mathbb{R}/\mathbb{Z}.$$

We may also choose γ so that $\gamma|_{[\tau_1, \tau_2]}$ (resp. $\gamma|_{[\tau_2, \tau_1]}$) is homotopic to some representative $\gamma_1 \colon [0, 1] \to \Sigma_1$ of w_1 (resp. some representative $\gamma_2 \colon [0, 1] \to \Sigma_2$ of w_2) relative to $\gamma_\star$, meaning that there is a homotopy between $\gamma|_{[\tau_1, \tau_2]}$ and γ_1 with endpoints (not necessarily fixed) in $\gamma_\star$. Here $[\tau_1, \tau_2] \subset \mathbb{R}/\mathbb{Z}$ is the interval linking τ_1 and τ_2 in the counterclockwise direction.

As $\gamma_{w_2 w_1}$ minimizes the quantity $i(\gamma, \gamma_\star)$ for $\gamma \in [\gamma_{w_2 w_1}]$ (see Lemma 2.1) we have either $i(\gamma_{w_2 w_1}, \gamma_\star) = 0$ or $i(\gamma_{w_2 w_1}, \gamma_\star) = 2$. If $i(\gamma_{w_2 w_1}, \gamma_\star) = 0$, then there exists a homotopy $H \colon [0, 1] \times \mathbb{R}/\mathbb{Z} \to \Sigma$ such that $H(0, \cdot) = \gamma$ and $H(1, \cdot) = \gamma$, so that $H(1, \tau) \notin \gamma_\star$ for any τ. As in the proof of Lemma 4.3, we may assume that H is transverse to $\gamma_\star$, in the sense that

$$dH(s, \tau)(T_{(s,\tau)}([0, 1] \times \mathbb{R}/\mathbb{Z})) + T_{H(s,\tau)}\gamma_\star = T_{H(s,\tau)}\Sigma \quad \text{for } H(s, \tau) \in \gamma_\star,$$

so that the preimage

$$H^{-1}(\gamma_\star) \subset [0, 1] \times \mathbb{R}/\mathbb{Z}$$

is an embedded submanifold of $[0, 1] \times \mathbb{R}/\mathbb{Z}$ (see Figure 5). As $H^{-1}(\gamma_\star) \cap \{s = 0\} = \{\tau_1, \tau_2\}$ and $H^{-1}(\gamma_\star) \cap \{s = 1\} = \varnothing$, it follows that there is an embedding $F \colon [0, 1] \to [0, 1] \times \mathbb{R}/\mathbb{Z}$ such that $F(0) = (0, \tau_1)$, $F(1) = (0, \tau_2)$ and

$$F(t) \in H^{-1}(\gamma_\star) \quad \text{for } t \in [0, 1].$$

As F is an embedding, F is homotopic (by a homotopy which preserves the endpoints) either to $J_{[\tau_1, \tau_2]}$ or to $J_{[\tau_2, \tau_1]}$, where $J_{[\tau, \tau']} \colon [0, 1] \to [0, 1] \times \mathbb{R}/\mathbb{Z}$ is the natural map that sends $[0, 1]$ to $\{0\} \times [\tau, \tau']$. We may assume without loss of generality that $F \sim J_{[\tau_1, \tau_2]}$. In particular, writing $F = (S, T)$, the map T is homotopic to $I_{[\tau_1, \tau_2]} = p_2 \circ J_{[\tau_1, \tau_2]}$, where $p_2 \colon [0, 1] \times \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ is the projection over the second factor. This means that there is $G \colon [0, 1] \times [0, 1] \to \mathbb{R}/\mathbb{Z}$ such that, for any $s, t \in [0, 1]$,

$$G(s, 0) = \tau_1, \quad G(s, 1) = \tau_2, \quad G(0, t) = \tau_1 + t(\tau_2 - \tau_1) \quad \text{and} \quad G(1, t) = T(t).$$
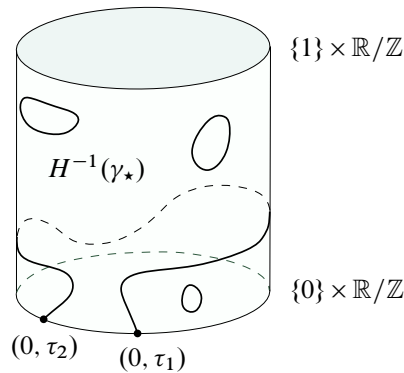
Figure 5: Proof of Lemma 4.6. The path linking $(0, \tau_1)$ to $(0, \tau_2)$ is the image of F.

Now we set $\widetilde{H}(s, t) = H(s S(t), G(s, t))$ for $s, t \in [0, 1]$. Then

$$\widetilde{H}(0, t) = \gamma(\tau_1 + t(\tau_2 - \tau_1)) \quad \text{and} \quad \widetilde{H}(1, t) = (H \circ F)(t) \qquad \text{for } t \in [0, 1],$$

$$\widetilde{H}(s, 0) = H(0, \tau_1) = x_1 \qquad \text{and} \quad \widetilde{H}(s, 1) = H(0, \tau_2) = x_2 \quad \text{for } s \in [0, 1].$$

We conclude that $t \mapsto \gamma|_{[\tau_1, \tau_2]}(\tau_1 + t(\tau_2 - \tau_1))$, and thus γ_1, is homotopic (relative to $\gamma_\star$) to some curve contained in $\gamma_\star$. Thus $w_1 = w_\star^k$, for some $k \in \mathbb{Z}$, in $\pi_1(\Sigma)$. As the inclusion $\pi_1(\Sigma_j) \to \pi_1(\Sigma)$ is injective by Lemma 4.4, the lemma follows. $\square$

Now, we need to understand when the geodesics given by $[w_2 w_1]$ and $[w_2' w_1']$ are the same. This is the purpose of the following:

Lemma 4.7 Take $w_j, w_j' \in \pi_1(\Sigma_j)$ for $j = 1, 2$ such that $i(\gamma_{[w_2 w_1]}, \gamma_\star) = 2$. Then $[w_2 w_1] = [w_2' w_1']$ as conjugacy classes of $\pi_1(\Sigma)$ if and only if there are $p, q \in \mathbb{Z}$ such that

(4-13) $$w_2 = w_{\star,2}^p w_2' w_{\star,2}^q \quad \text{and} \quad w_1 = w_{\star,1}^{-q} w_1' w_{\star,1}^{-p}.$$

Proof Again, let $\gamma \colon \mathbb{R}/\mathbb{Z} \to \Sigma$ be a smooth curve intersecting $\gamma_\star$ transversely such that

$$\{\tau \in \mathbb{R}/\mathbb{Z} : \gamma(\tau) \in \gamma_\star\} = \{\tau_1, \tau_2\} \quad \text{for some } \tau_1 \neq \tau_2 \in \mathbb{R}/\mathbb{Z},$$

with $\gamma([\tau_1, \tau_2]) \subset \Sigma_1$ and $\gamma([\tau_2, \tau_1]) \subset \Sigma_2$. Let $x_j = \gamma(\tau_j)$ for $j = 1, 2$, and chose arbitrary paths c_j contained in $\gamma_\star$ linking x_j to $x_\star$. Note that all the preceding choices can be made so that the curve $\gamma_1 = c_2 \gamma|_{[\tau_1, \tau_2]} c_1^{-1}$ (resp. $\gamma_2 = c_1 \gamma|_{[\tau_2, \tau_1]} c_2^{-1}$) represents $w_\star^p w_1 w_\star^q$ (resp. $w_\star^{-q} w_2 w_\star^{-p}$) for some $p, q \in \mathbb{Z}$. We may proceed in the same way to obtain γ', τ_1', τ_2', c_1', c_2', p' and q' so that the same properties hold with w_1 and w_2 replaced by w_1' and w_2'. By hypothesis, γ is freely homotopic to γ'. Thus we may find a smooth map $H \colon [0, 1] \times \mathbb{R}/\mathbb{Z} \to \Sigma$ such that $H(0, \cdot) = \gamma$ and $H(1, \cdot) = \gamma'$. As in Lemma 4.6, H may be chosen to be transverse to $\gamma_\star$, so that

$$H^{-1}(\gamma_\star) \subset [0, 1] \times \mathbb{R}/\mathbb{Z}$$

is a finite union of smooth embedded submanifolds of $[0, 1] \times \mathbb{R}/\mathbb{Z}$. Let $(x, \rho): \Sigma \to \mathbb{R}/\mathbb{Z} \times (-\varepsilon, \varepsilon)$ be coordinates near $\gamma_\star$ such that $\{\rho = 0\} = \gamma_\star$ and $|\rho| = \mathrm{dist}(\gamma_\star, \cdot)$, and such that $\{(-1)^{j-1}\rho \geqslant 0\} \subset \Sigma_j$. As $H^{-1}(\gamma_\star) \cap \{s = 0\} = \{\tau_1, \tau_2\}$ and $H^{-1}(\gamma_\star) \cap \{s = 1\} = \{\tau_1', \tau_2'\}$, we have two smooth embeddings $F_1, F_2: [0, 1] \to [0, 1] \times \mathbb{R}/\mathbb{Z}$ such that $F_j([0, 1]) \subset H^{-1}(\gamma_\star)$ and $F_j(0) = (0, \tau_j)$ for $j = 1, 2$, with $F_j(1) = \tau_1'$ or τ_2' (indeed we have $i(\gamma, \gamma_\star) = 2$ and thus there is a path in $H^{-1}(\gamma_\star)$ linking $\{s = 0\}$ to $\{s = 1\}$, since otherwise we could proceed as in the proof of Lemma 4.6 to obtain that $i(\gamma, \gamma_\star) = 0$). In fact, $F_1(1) = (1, \tau_1')$ and $F_2(1) = (1, \tau_2')$, which we shall prove later. Set $F_j = (S_j, T_j)$ and

$$\widetilde{H}(s, t) = H\big((1 - t)S_1(s) + tS_2(s), T_1(s) + t(T_2(s) - T_1(s))\big) \quad \text{for } s, t \in [0, 1].$$

Then

$$\widetilde{H}(0, t) = \gamma(\tau_1 + t(\tau_2 - \tau_1)) \quad \text{and} \quad \widetilde{H}(1, t) = \gamma'(\tau_1' + t(\tau_2' - \tau_1')) \quad \text{for } t \in [0, 1],$$
$$\widetilde{H}(s, 0) = H(S_1(s), T_1(s)) \quad \text{and} \quad \widetilde{H}(s, 1) = H(S_2(s), T_2(s)) \quad \text{for } s \in [0, 1].$$

For $j = 1, 2$, let $c_j(s), s \in [0, 1]$ be paths, contained in $\gamma_\star$ depending continuously on s and linking $T_j(s)$ to $x_\star$, such that $c_j(0) = c_j$. Then the construction of $\widetilde{H}$ shows that

$$c_2(0)\gamma|_{[\tau_1, \tau_2]}c_1(0)^{-1} \sim c_2(1)\gamma'|_{[\tau_1', \tau_2']}c_1(1)^{-1},$$

and reversing the role of τ_1 and τ_2 in the constructions made above,

$$c_1(0)\gamma|_{[\tau_2, \tau_1]}c_2(0)^{-1} \sim c_1(1)\gamma'|_{[\tau_2', \tau_1']}c_2(1)^{-1}.$$

Thus we obtain

$$w_\star^p w_1 w_\star^q = c_2(1)c_2'^{-1} w_\star^{p'} w_1' w_\star^{q'} c_1' c_1(1)^{-1} \quad \text{and} \quad w_\star^{-q} w_2 w_\star^{-p} = c_1(1)c_1'^{-1} w_\star^{-q'} w_2 w_\star^{-p'} c_2' c_2(1)^{-1},$$

which is the conclusion of Lemma 4.7 as the paths $c_1(1)c_1'^{-1}$ and $c_2(1)c_2'^{-1}$ are contained in $\gamma_\star$ (and, again, the inclusions $\pi_1(\Sigma_j) \to \pi_1(\Sigma)$ for $j = 1, 2$ are injective).

Thus it remains to show that $F_j(1) = (1, \tau_j')$ for $j = 1, 2$. We extend ρ into a smooth function $\rho: \Sigma \to \mathbb{R}$ such that $(-1)^{j-1}\rho > 0$ on $\Sigma_j \setminus \gamma_\star$. There exists a continuous path $G: [0, 1] \to ([0, 1] \times \mathbb{R}/\mathbb{Z}) \setminus H^{-1}(\gamma_\star)$ such that

$$G(0) \in \{0\} \times \,]\tau_1, \tau_2[\quad \text{and} \quad G(1) \in \{1\} \times (\mathbb{R}/\mathbb{Z} \setminus \{\tau_1', \tau_2'\}).$$

(Indeed, otherwise it would mean that there is a continuous path in $[0, 1] \times \mathbb{R}/\mathbb{Z}$ linking $(0, \tau_1)$ to $(0, \tau_2)$, which would imply, as in Lemma 4.6, that $i(\gamma, \gamma_\star) = 0$.) In particular, $\rho \circ H \circ G > 0$ since $\rho(H(0, \tau)) > 0$ for $\tau \in \,]\tau_1, \tau_2[$. Thus necessarily $G(1) \in \{1\} \times \,]\tau_1', \tau_2'[$, since $\rho(H(1, \tau)) < 0$ for $\tau \in \,]\tau_2', \tau_1'[$. Now, as $\mathrm{Im}(F_1) \cap \mathrm{Im}(F_2) = \varnothing$ (again, if the intersection was not empty we could find a path linking $(0, \tau_1)$ to $(0, \tau_2)$), we have that $G(1)$ lies in $]T_1(1), T_2(1)[$. Since $(\rho \circ H \circ G)(1) > 0$, it follows that $T_1(1) = \tau_1'$ and $T_2(1) = \tau_2'$. □

The above lemma motivates the next result:

Proposition 4.8 *There is a constant $C > 0$ such that the following holds. For any $w \in \pi_1(\Sigma_j)$ such that w is not a power of $w_{\star,j}$, there are $p_w, q_w \in \mathbb{Z}$ such that if $w' = w_{\star,j}^{p_w} w w_{\star,j}^{q_w}$,*

$$(4\text{-}14) \qquad \ell_\star(w_{\star,j}^p w' w_{\star,j}^q) \geq (|p| + |q|)\ell(\gamma_\star) + \ell_\star(w') - C \quad \text{for } p, q \in \mathbb{Z}.$$

In what follows, for any $x, y \in \widetilde{\Sigma}$ we will denote by $[x, y]$ the unique geodesic segment joining x and y. Before starting the proof of Proposition 4.8, we state a classical result valid in negatively curved spaces:

Lemma 4.9 *For each $\delta > 0$ there exists a constant $C > 0$ such that the following holds. For any sequence of geodesic segments $[x_0, x_1], [x_1, x_2], [x_2, x_3]$ in $\widetilde{\Sigma}$ such that $\text{dist}(x_1, x_2) \geq \delta$ and such that the angle between $[x_{j-1}, x_j]$ and $[x_j, x_{j+1}]$ is equal to $\pm\frac{1}{2}\pi$ for $j = 1, 2$,*

$$(4\text{-}15) \qquad \text{dist}(x_0, x_3) \geq \text{dist}(x_0, x_1) + \text{dist}(x_1, x_2) + \text{dist}(x_2, x_3) - C.$$

We will need the following intermediate result:

Fact 4.10 *For any $\varepsilon > 0$ there is $C > 0$ such that, for any pairwise distinct points $x, y, z \in \widetilde{\Sigma}$ such that the absolute value of the angle (taken in $]-\pi, \pi]$) between $[x, y]$ and $[y, z]$ is not smaller than ε, we have*

$$\text{dist}(x, z) \geq \text{dist}(x, y) + \text{dist}(y, z) - C.$$

Proof We prove the result by comparing $\widetilde{\Sigma}$ with a model space of constant curvature, as follows. Let $a = \text{dist}(x, y)$, $b = \text{dist}(y, z)$, $c = \text{dist}(x, z)$ and $\gamma = \angle([x, y], [y, z])$. Let $\widetilde{\Sigma}_k$ be a simply connected complete Riemannian surface with constant curvature $-k^2 < 0$ such that $\kappa \leq -k^2$ everywhere for some $k > 0$ (recall that κ is the curvature of Σ). Consider any points $\bar{x}, \bar{y}, \bar{z} \in \widetilde{\Sigma}_k$ such that

$$\text{dist}_k(\bar{x}, \bar{y}) = a, \quad \text{dist}_k(\bar{y}, \bar{z}) = b \quad \text{and} \quad \angle([\bar{x}, \bar{y}], [\bar{y}, \bar{z}]) = \gamma,$$

where dist_k is the distance in $\widetilde{\Sigma}_k$, and set $\bar{c} = \text{dist}_k(x, z)$. Then by a classical trigonometric formula for spaces of constant negative curvature (see [11, I.2.7]),

$$\text{ch}(k\bar{c}) = \text{ch}(ka)\,\text{ch}(kb) - \text{sh}(ka)\,\text{sh}(kb)\cos(\gamma).$$

As $\gamma \in]-\pi, \pi] \setminus]-\varepsilon, \varepsilon[$, we have $\cos(\gamma) \leq 1 - \eta$ for some $\eta \in]0, 1[$ depending on ε. Thus

$$\text{ch}(k\bar{c}) \geq \eta\,\text{ch}(ka)\,\text{ch}(kb).$$

Using $\frac{1}{2}\exp(t) \leq \text{ch}(t) \leq \exp(t)$ for $t \geq 0$, one gets

$$\bar{c} \geq a + b + \frac{\log\left(\frac{1}{4}\eta\right)}{k}.$$

As the scalar curvature of $\widetilde{\Sigma}$ is everywhere not greater than $-k^2$, the space $\widetilde{\Sigma}$ is a $\text{CAT}(-k^2)$ space; see [11, Theorem II.4.1]. In particular, by comparison, one obtains $c \geq \bar{c}$ (see [11, Proposition II.1.7]), which concludes the proof. $\qquad\square$
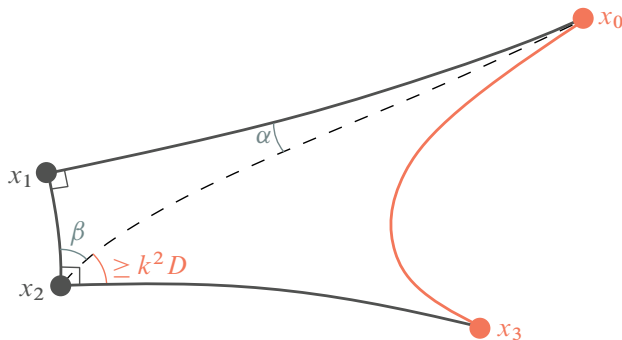
Figure 6: Proof of Lemma 4.9.

Proof of Lemma 4.9 Let x_0, x_1, x_2 and x_3 be as in the statement. For $j = 0, 1, 2$ we set $d_j = \text{dist}(x_j, x_{j+1})$. We first assume one of the numbers d_0 or d_2 is not greater than δ, say $d_0 \leqslant \delta$. Then Fact 4.10 (applied with $x = x_1$, $y = x_2$ and $z = x_3$) yields $\text{dist}(x_1, x_3) \geqslant d_1 + d_2 - C$, and thus

$$\text{dist}(x_0, x_3) \geqslant \text{dist}(x_1, x_3) - \text{dist}(x_0, x_1) \geqslant d_1 + d_2 + C - d_0 \geqslant d_0 + d_1 + d_2 + C - 2\delta.$$

Therefore we may assume that $d_0, d_2 \geqslant \delta$. Applying Fact 4.10 for the points x_0, x_1 and x_2 yields

$$(4\text{-}16) \qquad\qquad \text{dist}(x_0, x_2) \geqslant d_0 + d_1 - C.$$

For any pairwise distinct $x, y, z \in \widetilde{\Sigma}$, we denote by $\Delta(x, y, z)$ the triangle generated by x, y and z. Then as $d_0, d_1 \geqslant \delta$, the triangle $\Delta(x_0, x_1, x_2)$ contains some triangle $\Delta(x, y, z)$ with a right angle at y and $\text{dist}(x, y) = \text{dist}(y, z) = \delta$ (namely, $y = x_1$, $x \in [x_1, x_0]$ and $z \in [x_1, x_2]$). Clearly the area $|\Delta(x, y, z)|$ of $\Delta(x, y, z)$ is bounded from below by some constant $D > 0$ depending only on $\delta > 0$ (indeed, it suffices to verify this property for x, y and z lying in a compact set given by a finite union of fundamental domains of Σ). Therefore, $|\Delta(x_0, x_1, x_2)| \geqslant D$. Let α and β be the angles of $\Delta(x_0, x_1, x_2)$ at x_0 and x_2, respectively (see Figure 6). Let $\tilde{\mu}_g$ bet the Riemannian measure of $\widetilde{\Sigma}$, and $\tilde{\kappa}$ its scalar curvature. Then, by the Gauss–Bonnet formula [29, Theorem 9.3],

$$\int_{\Delta(x_0, x_1, x_2)} \tilde{\kappa} \, d\tilde{\mu}_g + \tfrac{1}{2}\pi + (\pi - \alpha) + (\pi - \beta) = 2\pi.$$

This gives

$$\beta \leqslant \tfrac{1}{2}\pi - \alpha - k^2 |\Delta(x_0, x_1, x_2)| \leqslant \tfrac{1}{2}\pi - k^2 D.$$

Therefore the angle between $[x_0, x_2]$ and $[x_2, x_3]$ is not smaller than $k^2 D$. In particular, we may apply Fact 4.10 to get $\text{dist}(x_0, x_3) \geqslant \text{dist}(x_0, x_2) + d_2 - C$ for some C depending only on $k^2 D$. Combining this with (4-16), we conclude the proof. $\qquad\square$

Proof of Proposition 4.8 We fix $j \in \{1, 2\}$ and write $w_\star = w_{\star, j}$ for simplicity. Let $w \in \pi_1(\Sigma_j)$ be such that $w \neq w_\star^k$ for any k. Then w is not the trivial element, and thus it is hyperbolic. Recall that $(\widetilde{\Sigma}, \tilde{g})$ is
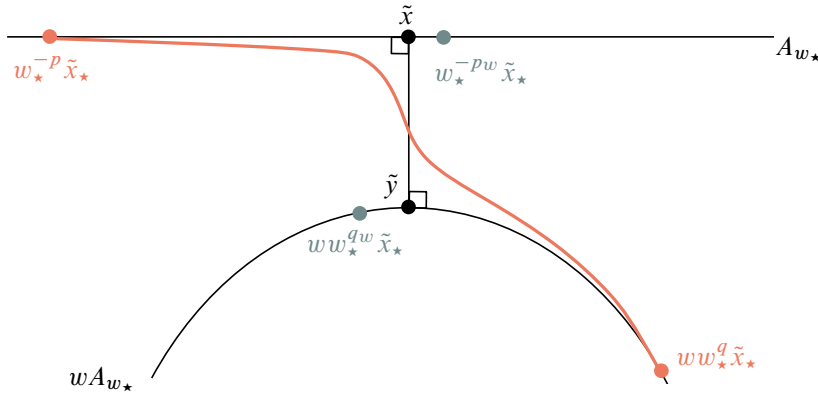
Figure 7: Proof of Proposition 4.8.

the universal cover of (Σ, g) and that $\pi_1(\Sigma)$ acts by deck transformations on $\widetilde{\Sigma}$. For any $u \in \pi_1(\Sigma) \setminus \{1\}$, we denote by

$$u_\pm = \lim_{k \to +\infty} u^{\pm k}(z)$$

the two distinct fixed points of u in the boundary at infinity $\partial_\infty \widetilde{\Sigma}$ of $\widetilde{\Sigma}$ (here z denotes any point in $\widetilde{\Sigma}$). We also denote by A_u the translation axis of u, that is, the unique complete geodesic of $(\widetilde{\Sigma}, \tilde{g})$ converging towards u_+ (resp. u_-) in the future (resp. in the past). Note that $A_{ww_\star w^{-1}} = wA_{w_\star}$. As the conjugacy classes $[ww_\star w^{-1}]$ and $[w_\star]$ both represent the geodesic $\gamma_\star$, we have either $A_{w_\star} = wA_{w_\star}$ or $A_{w_\star} \cap wA_{w_\star} = \varnothing$. Since w is not a power of $w_\star$, we necessarily have $A_{w_\star} \cap wA_{w_\star} = \varnothing$. Write $\gamma_\star = \{\varphi_s(z_\star) : s \in [0, \ell(\gamma_\star)]\}$ for some $z_\star = (x_\star, v_\star) \in M$. By hyperbolicity of the geodesic flow, there is $\delta > 0$ such that the following holds. For any $z \in M$ such that $\inf_{s \in \mathbb{R}} \text{dist}_M(z, \varphi_s(z_\star)) \leqslant \delta$,

$$(4\text{-}17) \qquad\qquad \varphi_{\ell(\gamma_\star)}(z) = z \implies z = \varphi_s(z_\star) \quad \text{for some } s \in \mathbb{R}.$$

As $\ell([ww_\star w^{-1}]) = \ell([w_\star]) = \ell(\gamma_\star)$, we obtain

$$(4\text{-}18) \qquad\qquad\qquad \text{dist}(A_{w_\star}, wA_{w_\star}) \geqslant \delta.$$

Let $\tilde{x} \in A_{w_\star}$ and $\tilde{y} \in wA_{w_\star}$ be the unique points such that $\text{dist}(\tilde{x}, \tilde{y}) = \text{dist}(A_{w_\star}, wA_{w_\star})$, and take $p, q \in \mathbb{Z}$. Then $\text{dist}(\tilde{x}, \tilde{y}) \geqslant \delta$ by (4-18), and thus we may apply Lemma 4.9 with the sequence of geodesic segments $[w_\star^{-p}\tilde{x}_\star, \tilde{x}]$, $[\tilde{x}, \tilde{y}]$, $[\tilde{y}, ww_\star^q\tilde{x}_\star]$ to obtain

$$\text{dist}(ww_\star^q\tilde{x}_\star, w_\star^{-p}\tilde{x}_\star) \geqslant \text{dist}(ww_\star^q\tilde{x}_\star, \tilde{y}) + \text{dist}(\tilde{y}, \tilde{x}) + \text{dist}(\tilde{x}, w_\star^{-p}\tilde{x}_\star) - C$$

for some $C > 0$ independent of w, p and q (see Figure 7). Next, let $p_w, q_w \in \mathbb{Z}$ such that

$$\text{dist}(\tilde{x}, w_\star^{-p_w}\tilde{x}_\star) < \ell(\gamma_\star) \quad \text{and} \quad \text{dist}(\tilde{y}, ww_\star^{q_w}\tilde{x}_\star) < \ell(\gamma_\star).$$

Then, for any $p, q \in \mathbb{Z}$,

$$\text{dist}(\tilde{x}, w_\star^{-p}\tilde{x}_\star) \geqslant |p - p_w|\ell(\gamma_\star) - \ell(\gamma_\star) \quad \text{and} \quad \text{dist}(\tilde{y}, ww_\star^q\tilde{x}_\star) \geqslant |q - q_w|\ell(\gamma_\star) - \ell(\gamma_\star),$$

which yields

$$\mathrm{dist}(w_\star^p w w_\star^q \tilde{x}_\star, \tilde{x}_\star) \geq (|p - p_w| + |q - q_w|)\ell(\gamma_\star) + \mathrm{dist}(\tilde{x}, \tilde{y}) - C - 2\ell(\gamma_\star).$$

Finally, we note that

$$\mathrm{dist}(\tilde{x}, \tilde{y}) \geq \mathrm{dist}(w w_\star^{q_w} \tilde{x}_\star, w_\star^{-p_w} \tilde{x}_\star) - 2\ell(\gamma_\star) = \ell_\star(w_\star^{p_w} w w_\star^{q_w}) - 2\ell(\gamma_\star). \qquad \square$$

Building on Lemmata 4.6 and 4.7 and Proposition 4.8, we prove Proposition 4.5:

Proof of Proposition 4.5 In what follows, C is a positive constant independent of L that may change at each line. First, assume that $h_1 = h_2 = h_\star$. For $j = 1, 2$ we denote by $\langle w_{\star,j} \rangle = \{w_{\star,j}^n : n \in \mathbb{Z}\}$ the infinite cyclic subgroup of $\pi_1(\Sigma_j)$ generated by $w_{\star,j}$, and we set $\pi_1(\Sigma_j)_\star = \pi_1(\Sigma_j) \setminus \langle w_{\star,j} \rangle$. Since $\ell_\star(w_{\star,j}^n) = |n|\ell(\gamma_\star)$, there is C such that, for any large L,

(4-19) $$C^{-1} e^{h_\star L} \leq N_{\star,j}(L) \leq C e^{h_\star L}$$

by (4-10), where $N_{\star,j}(L) = \#\{w \in \pi_1(\Sigma_j)_\star : \ell_\star(w) \leq L\}$. For $w \in \pi_1(\Sigma_j)_\star$, we set

$$\mathcal{C}_w = \{w_\star^p w w_\star^q : p, q \in \mathbb{Z}\} \subset \pi_1(\Sigma_j)_\star,$$

and we define $\mathscr{C}_j = \{\mathcal{C}_w : w \in \pi_1(\Sigma_j)_\star\}$. Note that the elements $\mathcal{C} \in \mathscr{C}_j$ are pairwise disjoint, and thus we have a partition $\bigsqcup_{\mathcal{C} \in \mathscr{C}_j} \mathcal{C}$ of $\pi_1(\Sigma_j)_\star$. We also write

$$\ell_\star(\mathcal{C}) = \inf\{\ell_\star(w) : w \in \mathcal{C}\} \quad \text{for } \mathcal{C} \in \mathscr{C}_j \text{ with } j = 1, 2.$$

Then Proposition 4.8 yields

$$\#\{w \in \mathcal{C} : \ell_\star(w) \leq L\} \leq C(L - \ell_\star(\mathcal{C}) + C)^2$$

for any $\mathcal{C} \in \mathscr{C}_j$ such that $\ell_\star(\mathcal{C}) \leq L$. Thus

$$N_{\star,j}(L) = \sum_{\substack{\mathcal{C} \in \mathscr{C}_j \\ \ell_\star(\mathcal{C}) \leq L}} \#\{w \in \mathcal{C} : \ell_\star(w) \leq L\} \leq C \sum_{\substack{\mathcal{C} \in \mathscr{C}_j \\ \ell_\star(\mathcal{C}) \leq L}} (L - \ell_\star(\mathcal{C}) + C)^2.$$

Let $\beta > 0$ be a large number. Then

(4-20) $$\sum_{\substack{\mathcal{C} \in \mathscr{C}_j \\ \ell_\star(\mathcal{C}) \leq L - \beta \log L}} (L - \ell_\star(\mathcal{C}) + C)^2 \leq (L + C)^2 \#\{\mathcal{C} \in \mathscr{C}_j : \ell_\star(\mathcal{C}) \leq L - \beta \log L\}.$$

However, using (4-19), we obtain

$$\#\{\mathcal{C} \in \mathscr{C}_j : \ell_\star(\mathcal{C}) \leq L - \beta \log L\} \leq N_{\star,j}(L - \beta \log L) \leq C L^{-h_\star \beta} e^{h_\star L}.$$

In particular, if $h_\star \beta > 2$, and if $A_\beta(L)$ denotes the left-hand side of (4-20), we have the bound $A_\beta(L) \ll N_{\star,j}(L)$ as $L \to \infty$. Thus, for large L,

$$C^{-1} N_{\star,j}(L) \leq \sum_{\substack{\mathcal{C} \in \mathscr{C}_j \\ \ell_\star(\mathcal{C}) \in [L - \beta \log L, L]}} (L - \ell_\star(\mathcal{C}) + C)^2 \leq (\beta \log L + C)^2 \#\{\mathcal{C} \in \mathscr{C}_j : \varepsilon L \leq \ell(\mathcal{C}) \leq L\},$$

where $\varepsilon > 0$ is any small number. This finally yields, for any large L,

$$(4\text{-}21) \qquad \#\{\mathcal{C} \in \mathscr{C}_j : \varepsilon L \leqslant \ell(\mathcal{C}) \leqslant L\} \geqslant \frac{C^{-1} e^{h_\star L}}{(\beta \log L + C)^2}.$$

For any $\mathcal{C} \in \mathscr{C}_j$, we choose some $w_\mathcal{C} \in \mathcal{C}$ such that $\ell_\star(w_\mathcal{C}) = \ell_\star(\mathcal{C})$. Then Lemmata 4.6 and 4.7 imply that we have a well-defined and injective map

$$\mathscr{C}_1 \times \mathscr{C}_2 \to \{\gamma \in \mathcal{P} : i(\gamma, \gamma_\star) = 2\}, \quad (\mathcal{C}_1, \mathcal{C}_2) \mapsto [w_{\mathcal{C}_2} w_{\mathcal{C}_1}] \equiv \gamma_{w_{\mathcal{C}_2} w_{\mathcal{C}_1}}.$$

Obviously, $\ell(\gamma_{w_2 w_1}) \leqslant \ell_\star(w_1) + \ell_\star(w_2)$ for any w_1 and w_2, and thus we get, for large L,

$$N(2, L) \geqslant \#\{(\mathcal{C}_1, \mathcal{C}_2) \in \mathscr{C}_1 \times \mathscr{C}_2 : \ell_\star(\mathcal{C}_1) + \ell_\star(\mathcal{C}_2) \leqslant L \text{ and } \ell_\star(\mathcal{C}_1), \ell_\star(\mathcal{C}_2) \geqslant \varepsilon L\}$$

$$\geqslant \sum_{\substack{\mathcal{C}_1 \in \mathscr{C}_1 \\ \varepsilon L \leqslant \ell_\star(\mathcal{C}_1) \leqslant L}} \#\{\mathcal{C}_2 \in \mathscr{C}_2 : \varepsilon L \leqslant \ell_\star(\mathcal{C}_2) \leqslant L - \ell_\star(\mathcal{C}_1)\} \geqslant \sum_{\substack{\mathcal{C}_1 \in \mathscr{C}_1 \\ \varepsilon L \leqslant \ell_\star(\mathcal{C}_1) \leqslant L}} \frac{C^{-1} e^{h_\star(L - \ell_\star(\mathcal{C}_1))}}{(\beta \log(L - \ell_\star(\mathcal{C}_1)) + C)^2}.$$

For simplicity, in what follows we will use the notation $f(\ell) = C^{-1} e^{h_\star \ell} / (\beta \log(\ell) + C)^2$ and $N(\mathscr{C}_1, L) = \#\{\mathcal{C} \in \mathscr{C}_j : \varepsilon L \leqslant \ell(\mathcal{C}) \leqslant L\}$. Fix some large number $\mu > 0$. Note that, if μ is large enough, there is $C > 0$ (depending on μ) such that, for any large ℓ,

$$(4\text{-}22) \qquad f(\ell + \mu) - f(\ell) \geqslant C^{-1} f(\ell).$$

There holds

$$(4\text{-}23) \quad N(2, L) \geqslant C^{-1} \sum_{k \in [\varepsilon L/\mu, L/\mu]} \big(N(\mathscr{C}_1, k\mu) - N(\mathscr{C}_1, (k-1)\mu)\big) f(L - (k-1)\mu)$$

$$\geqslant C^{-1} \sum_{k \in [\varepsilon L/\mu + 1, L/\mu - 1]} N(\mathscr{C}_1, k\mu) \big(f(L - (k-1)\mu) - f(L - k\mu)\big)$$

$$- N(\mathscr{C}_1, \varepsilon L + \mu) f(L - \varepsilon L),$$

where we used an Abel transformation in the last inequality. Next, note that by (4-19), one has $N(\mathscr{C}_1, L) \leqslant N_{\star,1}(L) \leqslant C e^{h_\star L}$. This yields

$$(4\text{-}24) \qquad N(\mathscr{C}_1, \varepsilon L + \mu) f(L - \varepsilon L) = \mathcal{O}(e^{h_\star L})$$

as $L \to \infty$. On the other hand, (4-22) gives, for any large L,

$$\sum_{k \in [\varepsilon L/\mu + 1, L/\mu - 1]} N(\mathscr{C}_1, k\mu) \big(f(L - (k-1)\mu) - f(L - k\mu)\big)$$

$$\geqslant \sum_{k \in [\varepsilon L/\mu + 1, L/\mu - 1]} N(\mathscr{C}_1, k\mu) f(L - k\mu)$$

$$\geqslant C^{-1} \sum_{k \in [\varepsilon L/\mu + 1, L/\mu - 1]} \frac{e^{h_\star k\mu}}{(\beta \log(k\mu) + C)^2} \frac{e^{h_\star(L - k\mu)}}{(\beta \log(L - k\mu) + C)^2} \geqslant \frac{C^{-1} L e^{h_\star L} (1 - \varepsilon)}{2\mu (\log(L) + C)^4}.$$

We conclude the proof of Proposition 4.5 for the case $h_1 = h_2$ by combining this last estimate with (4-23) and (4-24).

If $h_1 \neq h_2$, say $h_1 > h_2$ (the case $h_1 < h_2$ is identical), one is able to obtain the desired bound by considering, for example, the injective map $\mathscr{C}_1 \to \{\gamma \in \mathcal{P} : i(\gamma, \gamma_\star) = 2\}$ given by $\mathcal{C} \mapsto [a_g w_{\mathcal{C}}]$ and by using (4-21). $\qquad \square$

4.2.2 Upper bound Clearly, each $\gamma \in \mathcal{P}_2$ with $\ell(\gamma) \leqslant L$ may be represented by the conjugacy class of $w_1 w_2$ for some $w_j \in \pi_1(\Sigma_j)$ with $\ell_\star(w_1) + \ell_\star(w_2) \leqslant L + C$. Therefore, (4-5) implies

$$N(2, L) \leqslant \#\{(w_1, w_2) \in \pi_1(\Sigma_1) \times \pi_1(\Sigma_2) : \ell_\star(w_1) + \ell_\star(w_2) \leqslant L + C\}$$
$$\leqslant \sum_{k=0}^{L+C} C \exp(h_1 k) \exp(h_2(L - k + C)),$$

which gives, for large L, if $h_\star = \max(h_1, h_2)$,

$$N(2, L) \leqslant \begin{cases} CL \exp(h_\star L) & \text{if } h_1 = h_2, \\ C \exp(h_\star L) & \text{if } h_1 \neq h_2. \end{cases}$$

Iterating this process we obtain (with C depending on n)

$$N(2n, L) \leqslant \begin{cases} CL^{2n-1} \exp(h_\star L) & \text{if } h_1 = h_2, \\ CL^{n-1} \exp(h_\star L) & \text{if } h_1 \neq h_2. \end{cases}$$

4.3 Relative growth of closed geodesics with a small intersection angle

For $x = \gamma_\star(\tau) \in \mathrm{Im}(\gamma_\star)$, we let $v_\star(x) = \dot\gamma_\star(\tau)$. For any $\eta > 0$ small, we consider the number $N(n, \eta, L) = \#\mathcal{P}_{\eta,n}(L)$, where $\mathcal{P}_{\eta,n}(L)$ is the set of closed geodesics $\gamma \colon \mathbb{R}/\ell(\gamma)\mathbb{Z} \to \Sigma$ of length not greater than L, intersecting $\gamma_\star$ exactly n times, and such that there is $t \in \mathbb{R}/\ell(\gamma)\mathbb{Z}$ with $\gamma(t) \in \mathrm{Im}(\gamma_\star)$ and

$$\mathrm{angle}\big(\dot\gamma(t), v_\star(\gamma(t))\big) < \eta \quad \text{or} \quad \mathrm{angle}\big(\dot\gamma(t), -v_\star(\gamma(t))\big) < \eta.$$

The purpose of this section is to prove the following estimate:

Lemma 4.11 *Let $n \geqslant 1$. For any $\varepsilon, L_0 > 0$, there exists $\eta_0 > 0$ such that, for any $\eta \in {]0, \eta_0[}$ and any large L,*

$$(4\text{-}25) \qquad N(1, \eta, L) \leqslant 4N(1, L - L_0) \quad \text{and} \quad N(n, \eta, L) \leqslant \varepsilon L^{n-1} \exp(h_\star L)$$

if $\gamma_\star$ is not separating, and

$$(4\text{-}26) \quad N(2, \eta, L) \leqslant 4N(2, L - L_0) \quad \text{and} \quad N(2n, \eta, L) \leqslant \begin{cases} \varepsilon L^{2n-1} \exp(h_\star L) & \text{if } h_1 = h_2, \\ \varepsilon L^{n-1} \exp(h_\star L) & \text{if } h_1 \neq h_2, \end{cases}$$

if $\gamma_\star$ is separating.

Proof We first prove the lemma when $\gamma_\star$ is assumed not separating. Let $\gamma \colon [0, \ell(\gamma)] \to \Sigma$ be an element of $\mathcal{P}_{\eta,n}(L)$ parametrized by arc length. Let $0 \leqslant t_1 < t_2 < \cdots < t_n < \ell(\gamma)$ be such that $\gamma(t_j) \in \mathrm{Im}(\gamma_\star)$. For every $j = 1, \ldots, n$, we choose a path c_j contained in $\mathrm{Im}(\gamma_\star)$ of length not greater than $\ell(\gamma_\star)$ that links

$x_j = \gamma(t_j)$ to $x_\star$. Recall that we have a map $q_\star \colon \Sigma_\star \to \Sigma$ given by the identification of the boundary components of $\Sigma_\star$. Write $q_\star^{-1}(x_\star) = \{x_\star, \bar{x}_\star\}$, where we chose some $x_\star \in \Sigma_\star$ with $q_\star(x_\star) = x_\star$, as in Section 4.1. Then γ is freely homotopic to the composition

$$w_1 w_2 \cdots w_n, \quad \text{where } w_j = c_{j+1}\gamma|_{[t_j, t_{j+1}]}c_j^{-1} \in \pi_1(\Sigma) \text{ for } j = 1, \dots, n,$$

with the convention that $t_{n+1} = \ell(\gamma)$ and $c_{n+1} = c_1$. Note also that

$$\ell_\star(w_j) \leqslant |t_{j+1} - t_j| + 2\ell(\gamma_\star).$$

In fact, the elements w_j actually define elements of the space $\pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\})$, that is, the space of equivalence classes of paths $c \colon [0, 1] \to \Sigma_\star$ with $c(0), c(1) \in \{x_\star, \bar{x}_\star\}$, where two paths are equivalent if they are homotopic via a homotopy preserving the endpoints. The space $\pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\})$ is not a group (we may not be able to concatenate two paths); however, we have a natural map $\pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\}) \to \pi_1(\Sigma)$. In particular, for any $u_1, \dots, u_n \in \pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\})$, the composition $u_n \cdots u_1$ is well defined in $\pi_1(\Sigma)$. For any $u \in \pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\})$, we will denote by $\ell_\star(u)$ the infimum of the lengths of curves in the equivalence class u.

Up to reparametrizing of γ, we may assume that $t_1 = 0$, and either $\angle(v, v_\star) < \eta$ or $\angle(v, -v_\star) < \eta$, where we set $x = \gamma(0)$, $v_\star = v_\star(x)$ and $v = \dot{\gamma}(0)$. We will first assume that $\angle(v, v_\star) < \eta$. Let $L_0 > 0$ be a large number and $\varepsilon > 0$ be small. By continuity of the geodesic flow (φ_t), there is $\eta_0 > 0$ such that, if $\eta < \eta_0$,

$$\operatorname{dist}_M(\varphi_t(v), \varphi_t(v_\star)) \leqslant \varepsilon \quad \text{for } t \in [0, L_0].$$

Let K be a positive integer such that $K \in [L_0/\ell(\gamma_\star) - 1, L_0/\ell(\gamma_\star)]$, so that

$$\operatorname{dist}_\Sigma\big(\pi(\varphi_{K\ell(\gamma_\star)}(v)), x\big) < \varepsilon.$$

Let c_K be a path in Σ of length not greater than ε linking $\pi(\varphi_{K\ell(\gamma_\star)}(v))$ and x. Then, if $\varepsilon > 0$ is small enough,[11]

$$c_1 c_K \gamma|_{[0, K\ell(\gamma_\star)]}c_1^{-1} = a_g^K \quad \text{in } \pi_1(\Sigma).$$

In particular, $w_1 = w_1' a_g^K$ in $\pi_1(\Sigma)$, where $w_1' = c_2\gamma|_{[K\ell(\gamma_\star), t_2]}c_K^{-1}c_1^{-1}$. Note also that

$$\ell_\star(w_1') \leqslant |t_2 - K\ell(\gamma_\star)| + 2\ell(\gamma_\star) + \varepsilon,$$

where w_1' is seen as an element of $\pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\})$. Note that if we had assumed $\angle(v, -v_\star) < \eta$, we would have obtained the same factorization with a_g^{-K} instead of a_g^K. Next, let

$$A_{K,n}(L) = \left\{ (w_1, \dots, w_n) \in \pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\})^n : \sum_{j=1}^n \ell_\star(w_j) \leqslant L + (2n - K)\ell(\gamma_\star) + \varepsilon \right\},$$

---

[11]If $\varepsilon > 0$ is small enough, we have the following property. For any $x \in \Sigma$ and $L > 0$, if we are given two paths $c, c' \colon [0, L] \to \Sigma$ such that $c(0) = c'(0) = c(L) = c'(L) = x$ and $\operatorname{dist}_\Sigma(c(t), c'(t)) < \varepsilon$, then c and c' define the same element in $\pi_1(\Sigma, x)$.

and consider the map $\Psi_{K,n,\pm} : A_{K,n}(L) \to \mathcal{P}$ given by $(w_1, \ldots, w_n) \mapsto [w_1 \cdots w_n a_{\mathrm{g}}^{\pm K}]$. Then the discussion above shows that

$$\mathcal{P}_{\eta,n}(L) \subset \mathrm{Im}(\Psi_{K,n,+}) \cup \mathrm{Im}(\Psi_{K,n,-}).$$

In particular, $N(n, \eta, L) \leq 2 \# A_{K,n}(L)$. Next, we obtain a bound on $A_{K,n}(L)$ as follows. Let $c_\star$ be a path connecting $\bar{x}_\star$ and $x_\star$ in $\Sigma_\star$, so that the image of $c_\star^{-1}$ in $\pi_1(\Sigma)$ is b_{g} (see Figure 2). Then it is not hard to see that, for any $w \in \pi_1(\Sigma_\star, \{x_\star, \bar{x}_\star\})$, there is $u \in \pi_1(\Sigma_\star, x_\star)$ such that w can be written as

$$u, \quad c_\star u, \quad u c_\star^{-1} \quad \text{or} \quad c_\star u c_\star^{-1}$$

(depending on the endpoints of w), with $\ell_\star(u) \leq \ell_\star(w) + 2\ell(c_\star)$. This immediately gives

$$\# A_{K,1}(L) \leq 4 \,\#\{u \in \pi_1(\Sigma_\star) : \ell_\star(u) \leq L\} \leq C \exp(h_\star L).$$

As in Section 4.1.2, we obtain, for some $C_n > 0$ depending only n,

$$\# A_{K,n}(L) \leq C_n L^{n-1} \exp(h_\star(L - L_0)),$$

where we used that $K\ell(\gamma_\star) \geq L_0 - \ell(\gamma_\star)$. This proves the second part of (4-25). For the first part, we proceed as follows. With the notation of the proof of Proposition 4.5, one has well-defined maps

$$\Psi_{K,1,\pm,r}, \Psi_{K,1,\pm,l} : \{\mathcal{C} \in \mathscr{C} : \ell_\star(w) \leq L - K\ell(\gamma_\star)\} \to \{\gamma \in \mathcal{P}_1 : \ell(\gamma) \leq L + 2C\},$$

given respectively by $\mathcal{C} \mapsto [a_{\mathrm{g}}^{\pm K} w b_{\mathrm{g}}]$ and $\mathcal{C} \mapsto [b_{\mathrm{g}}^{-1} w a_{\mathrm{g}}^{\pm K}]$, where w is any element of $\mathcal{C}$. Next, we remark that the above discussion implies that every $\gamma \in \mathcal{P}_{\eta,1}(L)$ can be written as

$$[a_{\mathrm{g}}^{\pm K} w b_{\mathrm{g}}] \quad \text{or} \quad [b_{\mathrm{g}}^{-1} w a_{\mathrm{g}}^{\pm K}]$$

for some $w \in \pi_1(\Sigma_\star)$ with $\ell_\star(w) \leq L - K\ell(\gamma_\star) + C$. Therefore the union of the images of the maps $\Psi_{K,1,\pm,r}$ and $\Psi_{K,1,\pm,l}$ contains $\mathcal{P}_\eta(L + 2C)$, and thus

$$N(1, \eta, L) \leq 4 \,\#\{\mathcal{C} \in \mathscr{C} : \ell_\star(w) \leq L - K\ell(\gamma_\star) + 2C\} \leq 4N(1, L - K\ell(\gamma_\star) + 3C),$$

where we used the first inequality of (4-8). This gives the first part of (4-25).

Next, assume that $\gamma_\star$ is separating. Then, as above, every $\gamma : [0, \ell(\gamma)] \to \Sigma$ such that $\gamma \in \mathcal{P}_{2n,\eta}(L)$ can be written as a composition $w_{1,1} w_{1,2} \cdots w_{1,n} w_{2,n}$ for some $w_{k,j} \in \pi_1(\Sigma_k)$ for $k = 1, 2$ and $j = 1, 2, \ldots, n$, with

$$\sum_{j=1}^{n} \ell_\star(w_{2,j}) + \ell_\star(w_{1,j}) \leq \ell(\gamma) + 4n\ell(\gamma_\star).$$

Now, if η is small, we may proceed as before to obtain (up to reparametrization of γ) that $w_{1,1} = w_{\star,1}^{\pm K} w_{1,1}'$ or $w_{1,1} = w_{1,1}' w_{\star,1}^{\pm K}$ for some $w_{1,1}' \in \pi_1(\Sigma_1)$ with

$$\ell_\star(w_{1,1}') \leq \ell_\star(w_{1,1}) - K\ell(\gamma_\star) + C.$$

Here K is a large number depending on η (ie such that $K \to \infty$ as $\eta \to 0$) and $C > 0$ is a constant independent of γ and K. Thus we get

$$N(2n, \eta, L)$$
$$\leqslant C \, \# \Bigg\{ (w_{1,1}, w_{2,1}, \dots, w_{1,n}, w_{2,n}) : w_{k,j} \in \pi_1(\Sigma_k), \sum_{j=1}^{n} \ell_\star(w_{1,j}) + \ell_\star(w_{2,j}) \leqslant L - K\ell(\gamma_\star) + Cn \Bigg\}.$$

Then we obtain the second part of (4-26) by proceeding as in Section 4.2.2. For the first part of (4-26), we proceed as follows. For $w_j \in \pi_1(\Sigma_j)_\star$, we define

$$\mathcal{C}_{w_1, w_2} = \{ (w_1', w_2') : [w_1' w_2'] = [w_1 w_2] \}$$

and $\ell_\star(\mathcal{C}_{w_1, w_2}) = \inf\{ \ell_\star(w_1') + \ell_\star(w_2') : (w_1', w_2') \in \mathcal{C}_{(w_1, w_2)} \}$. We also introduce the notation $\mathscr{C}_{1,2} = \{ \mathcal{C}_{w_1, w_2} : w_j \in \pi_1(\Sigma_j)_\star \}$. By Lemmata 4.6 and 4.7, we have well-defined maps

$$\Psi_{K,1,\pm,r}, \Psi_{K,1,\pm,l} : \{ \mathcal{C} \in \mathscr{C}_{1,2} : \ell_\star(\mathcal{C}_{w_1, w_2}) \leqslant L - K\ell(\gamma_\star) \} \to \{ \gamma \in \mathcal{P}_2 : \ell(\gamma) \leqslant L \}$$

given respectively by $\mathcal{C} \mapsto [w_1 w_{\star,1}^{\pm K} w_2]$ and $\mathcal{C} \mapsto [w_{\star,1}^{\pm K} w_1 w_2]$. By the discussion above, the union of the images of those maps contains $\mathcal{P}_{2,\eta}(L)$. Therefore

$$N(2, \eta, L) \leqslant 4 \, \#\{ \mathcal{C} \in \mathscr{C}_{1,2} : \ell_\star(\mathcal{C}_{w_1, w_2}) \leqslant L - K\ell(\gamma_\star) \} \leqslant 4N(2, L - K\ell(\gamma_\star)),$$

where we used Lemmata 4.6 and 4.7 again in the last inequality. The first part of (4-26) follows. □

5 A Tauberian argument

The goal of this section is to give an asymptotic growth of the quantity

$$N_\pm(n, \chi, t) = \sum_{\substack{\gamma \in \mathcal{P} \\ i(\gamma_\star, \gamma) = n \\ \ell(\gamma) \leqslant t}} I_{\star, \pm}(\gamma, \chi)$$

as $t \to +\infty$, where $\chi \in C_c^\infty(\partial \setminus \partial_0)$ and $I_{\star, \pm}(\gamma, \chi) = \prod_{z \in I_{\star, \pm}(\gamma)} \chi^2(z)$.

5.1 The case $\gamma_\star$ is not separating

By [15, Theorem 3 and Section 6.2], the zeta function

$$\zeta_{\Sigma_\star}(s) = \prod_{\gamma \in \mathcal{P}_\star} (1 - e^{-s\ell(\gamma)})$$

extends meromorphically to the whole complex plane, and moreover we may write

$$\frac{\zeta'_{\Sigma_\star}(s)}{\zeta_{\Sigma_\star}(s)} = \sum_{k=0}^{2} (-1)^k \operatorname{tr}^\flat (e^{\pm \varepsilon s} \varphi_{\mp \varepsilon}^* R_{\pm, \delta}(s)|_{\Omega_c^k(M_\delta) \cap \ker \iota_X}),$$

where the flat trace is computed on M_δ. Here $\mathcal{P}_\star$ denotes the set of primitive closed geodesics of $(\Sigma_\star, g)$. By [12], we may apply [35, Proposition 9] (see also [36, Theorem 9.1]) to obtain that $\zeta_{\Sigma_\star}$ is holomorphic

in $\{\operatorname{Re}(s) \geqslant h_\star\}$, except for a simple pole at $s = h_\star$, where $h_\star > 0$ is the topological entropy of the geodesic flow of $(\Sigma_\star, g)$ restricted to its trapped set. Write the Laurent expansion given in [Section 2.6](#) of $R_{\pm,\delta}(s)$ near $s = h_\star$ as

$$R_{\pm,\delta}(s) = Y_{\pm,\delta}(s) + \frac{\Pi_{\pm,\delta}(h_\star)}{s - h_\star} + \sum_{j=2}^{J(h_\star)} \frac{(X \pm h_\star)^{j-1} \Pi_{\pm,\delta}(h_\star)}{(s - h_\star)^j} : \Omega_c^\bullet(M_\delta) \to \mathcal{D}'^\bullet(M_\delta).$$

By [15, (5.8)], we have $\operatorname{tr}^\flat(e^{\pm \varepsilon h_\star} \varphi_{\mp\varepsilon}^* \Pi_{\pm,\delta}(h_\star)) = \operatorname{rank} \Pi_{\pm,\delta}(h_\star)$ and

$$\operatorname{tr}^\flat(\varphi_{\mp\varepsilon}^* (X \pm h_\star)^j \Pi_{\pm,\delta}(h_\star)) = 0 \quad \text{for } j = 1, \ldots, J(h_\star) - 1.$$

We write $\Omega^k = \Omega_c^k(M_\delta)$ and $\Omega_0^k = \Omega^k \cap \ker \iota_X$. Then, by [18, Propositions 2.4 and 4.4], the map $s \mapsto R_{\pm,\delta}(s)|_{\Omega_0^0}$ has no pole in $\{\operatorname{Re}(s) > 0\}$. Since $\Omega_0^2 = \Omega_0^0 \wedge d\alpha$, and $R_{\pm,\delta}(s)|_{\Omega_0^2} = R_{\pm,\delta}(s)|_{\Omega_0^0} \wedge d\alpha$ (because $\varphi_t^* \alpha = \alpha$), it follows that $s \mapsto R_{\pm,\delta}(s)|_{\Omega_0^2}$ has no poles in $\{\operatorname{Re}(s) > 0\}$. In particular, the residue of $\zeta'_{\Sigma_\star}(s)/\zeta_{\Sigma_\star}(s)$ at $s = h_\star$ is given by $\operatorname{rank}(\Pi_{\pm,\delta}(h_\star)|_{\Omega_0^1})$, and since $\zeta_{\Sigma_\star}(s)$ has a simple pole at $s = h_\star$, this residue is equal to 1. Therefore,

$$\operatorname{rank}(\Pi_{\pm,\delta}(h_\star)|_{\Omega_0^1}) = 1.$$

In particular, $(X \pm h_\star)^j \Pi_{\pm,\delta} = 0$ for each $j = 1, \ldots, J(h_\star) - 1$. As $R_{\pm,\delta}(s)$ commutes with ι_X, it preserves the spaces Ω_0^k. Writing $\Omega^k = \Omega_0^k \oplus \alpha \wedge \Omega_0^{k-1}$ we have, for any $w = u + \alpha \wedge v$ with $\iota_X u = 0$ and $\iota_X v = 0$,

$$\Pi_{\pm,\delta}(h_\star)|_{\Omega^2}(u + \alpha \wedge v) = \Pi_{\pm,\delta}(h_\star)|_{\Omega_0^2}(u) + \alpha \wedge \Pi_{\pm,\delta}(h_\star)|_{\Omega_0^1}(v).$$

Thus $\Pi_{\pm,\delta}(h_\star)|_{\Omega^2} = \alpha \wedge \iota_X \Pi_{\pm,\delta}(h_\star)|_{\Omega_0^1}$. By [Proposition 3.2](#) and the fact that $\varphi_{\pm\varepsilon}^* \Pi_{\pm,\delta}(h_\star) = e^{\pm \varepsilon h_\star} \Pi_{\pm,\delta}(h_\star)$, we have, near $s = h_\star$,

$$(5\text{-}1) \qquad \chi \widetilde{S}_\pm(s) \chi = \chi Y_\pm(s) \chi + \frac{\chi \psi^* \iota^* \iota_X \Pi_{\pm,\delta}(h_\star) \iota_* \chi}{s - h_\star},$$

where $s \mapsto Y_\pm(s)$ is holomorphic in a neighborhood of $h_\star$. We write

$$\Pi_{\pm,\partial} = \psi^* \iota^* \iota_X \Pi_{\pm,\delta}(h_\star) \iota_* : \Omega^\bullet(\partial) \to \mathcal{D}'^\bullet(\partial).$$

Then, by what precedes, and since $\iota_X \Pi_{\pm,\delta}(h_\star)|_{\Omega^1} = 0$, we obtain that $\operatorname{rank}(\Pi_{\pm,\partial}) \leqslant 1$. Finally, for any $\chi \in C_c^\infty(\partial \setminus \partial_0)$, we set

$$c_\pm(\chi) = \operatorname{tr}_s^\flat(\chi \Pi_{\pm,\partial} \chi).$$

Lemma 5.1 *Let $\chi \in C_c^\infty(\partial \setminus \partial_0)$ be such that $c_\pm(\chi) > 0$. Then*

$$N_\pm(n, \chi, t) \sim \frac{(c_\pm(\chi) t)^n}{n!} \frac{e^{h_\star t}}{h_\star t} \quad \text{as } t \to +\infty.$$

Proof Because $\chi \Pi_{\pm,\partial}$ is of rank one, it follows that $\operatorname{tr}_s^\flat((\chi \Pi_{\pm,\partial})^n) = c_\pm(\chi)^n$ for any $n \geqslant 1$ (since the flat trace of a finite-rank operator coincides with its usual trace), and thus

$$\operatorname{tr}_s^\flat((\chi \widetilde{S}_\pm(s) \chi)^n) = \frac{c_\pm(\chi)^n}{(s - h_\star)^n} + \mathcal{O}((s - h_\star)^{-n+1}) \quad \text{as } s \to h_\star.$$

Note that here we implicitly used the fact that the flat trace of products of the form

$$(5\text{-}2) \qquad (\chi Y_\pm(s)\chi)^{k_1}(\chi\Pi_{\pm,\partial}\chi)^{\ell_1}(\chi Y_\pm(s)\chi)^{k_2}(\chi\Pi_{\pm,\partial}\chi)^{\ell_2}\cdots$$

makes sense. Indeed, note that both $\mathrm{WF}(\chi\Pi_{\pm,\partial}\chi)$ and $\mathrm{WF}(\chi Y_\pm(s)\chi)$ are contained in $\mathrm{WF}(\chi\widetilde{\mathcal{S}}_\pm(s)\chi)$ by (5-1) and Cauchy's integral formula. Thus we may reproduce the proofs of Lemma 3.5, Remark 3.6 and Proposition 3.7 to obtain that the composition (5-2) is well defined and that its flat trace makes sense. Next, set $\eta_{n,\chi}(s) = \mathrm{tr}_s^\flat\big((\chi\widetilde{\mathcal{S}}_\pm(s)\chi)^n\big)$ and

$$g_{n,\chi}(t) = \sum_{\substack{\gamma\in\mathcal{P}\\ i(\gamma,\gamma_\star)=n}} \ell^\#(\gamma) \sum_{\substack{k\geqslant 1\\ k\ell(\gamma)\leqslant t}} I_{\star,\pm}(\gamma,\chi)^k \quad \text{for } t\geqslant 0.$$

Now, if $G_{n,\chi}(s) = \int_0^{+\infty} g_{n,\chi}(t)e^{-ts}\,dt$, a simple computation leads to

$$G_{n,\chi}(s) = \frac{1}{s}\sum_{i(\gamma,\gamma_\star)=n} \ell^\#(\gamma)e^{-s\ell(\gamma)}I_{\star,\pm}(\gamma,\chi)^{\ell(\gamma)/\ell^\#(\gamma)} = -\frac{\eta'_{n,\chi}(s)}{ns},$$

where the last equality comes from Proposition 3.7. Using the expansion

$$\eta'_{n,\chi}(s) = -nc_\pm(\chi)^n(s-h_\star)^{-(n+1)} + \mathcal{O}((s-h_\star)^{-n}) \quad \text{as } s\to h_\star,$$

we obtain

$$G_{n,\chi}(h_\star s) = \frac{c_\pm(\chi)^n}{h_\star^{n+2}(s-1)^{n+1}} + \mathcal{O}((s-h_\star)^{-n}) \quad \text{as } s\to h_\star.$$

Then, applying the Tauberian theorem of Delange [14, théorème III],

$$\frac{1}{h_\star}g_{n,\chi}\left(\frac{t}{h_\star}\right) \sim \frac{c_\pm(\chi)^n}{h_\star^{n+2}}\frac{e^t}{n!}t^n \quad \text{as } t\to +\infty,$$

and so

$$(5\text{-}3) \qquad g_{n,\chi}(t) \sim \frac{(c_\pm(\chi)t)^n}{n!h_\star}\exp(h_\star t).$$

Now note that, if $\mathcal{P}_n$ is the set of primitive closed geodesics γ with $i(\gamma,\gamma_\star)=n$,

$$g_{n,\chi}(t) \leqslant \sum_{\substack{\gamma\in\mathcal{P}_n\\ \ell(\gamma)\leqslant t}} \ell(\gamma)\left\lfloor\frac{t}{\ell(\gamma)}\right\rfloor I_{\star,\pm}(\gamma,\chi) \leqslant tN(n,\chi,t).$$

As a consequence,

$$(5\text{-}4) \qquad \liminf_{t\to +\infty} N_\pm(n,\chi,t)\frac{n!h_\star t}{(c_\pm(\chi)t)^n e^{h_\star t}} \geqslant 1.$$

For the other bound, we use the a priori bound, obtained in Section 4.1.2,

$$(5\text{-}5) \qquad N_\pm(n,\chi,t) \leqslant N(n,t) \leqslant \frac{Ct^n}{n!}\frac{e^{h_\star t}}{h_\star t}$$

to deduce that, for any $\sigma > 1$,

$$(5\text{-}6) \qquad \limsup_{t\to +\infty} N_\pm\left(n,\chi,\frac{t}{\sigma}\right)\frac{n!}{t^n}\frac{h_\star t}{e^{h_\star t}} = 0.$$

Now we may write

$$(5\text{-}7) \qquad N_\pm(n, \chi, t) = N_\pm\left(n, \chi, \frac{t}{\sigma}\right) + \sum_{\substack{\gamma \in \mathcal{P} \\ i(\gamma_\star, \gamma) = n \\ t/\sigma \leqslant \ell(\gamma) \leqslant t}} I_{\star, \pm}(\gamma, \chi)$$

$$\leqslant N_\pm\left(n, \chi, \frac{t}{\sigma}\right) + \frac{\sigma}{t} \sum_{\substack{\gamma \in \mathcal{P} \\ i(\gamma_\star, \gamma) = n \\ t/\sigma \leqslant \ell(\gamma) \leqslant t}} I_{\star, \pm}(\gamma, \chi)\ell(\gamma) \leqslant N_\pm\left(n, \chi, \frac{t}{\sigma}\right) + \frac{\sigma}{t} g_{n, \chi}(t),$$

which gives, with (5-3) and (5-6),

$$\limsup_{t \to +\infty} N_\pm(n, \chi, t) \frac{n!}{(c_\pm(\chi)t)^n} \frac{h_\star t}{e^{h_\star t}} \leqslant \sigma.$$

As $\sigma > 1$ is arbitrary, the lemma is proven. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

Remark 5.2 If we assume that $c_\pm(\chi) = 0$, then with the notation of the above proof, the map $s \mapsto \eta_{1, \chi}(s)$ has no pole on the line $\{\mathrm{Re}(s) = h_\star\}$. In particular, we may reproduce the arguments of the aforementioned proof, replacing $g_{n, \chi}(t)$ by $g_{n, \chi}(t) + \exp(h_\star t)$, to obtain that $s \mapsto \int_0^\infty (g_{n, \chi}(t) + \exp(h_\star t)) \exp(-ts)\, dt$ has a pole of order 1 at $s = h_\star$, which implies that $g_{n, \chi}(t) + \exp(h_\star t) \sim \exp(h_\star t)$ as $t \to \infty$. This gives $g_{n, \chi}(t) \ll_{t \to \infty} \exp(h_\star t)$, and hence

$$N_\pm(1, \chi, t) \ll \frac{\exp(h_\star t)}{t} \quad \text{as } t \to \infty,$$

where we used the last line of (5-7) and (5-5). Note that this bound is incompatible with the one provided by Proposition 4.2; this will help us to prove that $c_\pm(\chi) > 0$, by showing that $N(1, t)$ can be controlled by $N_\pm(1, \chi, t)$ whenever χ has enough support (see Section 6.1).

5.2 The case $\gamma_\star$ is separating

In this case, Σ_δ consists of two surfaces, $\Sigma_\delta^{(1)}$ and $\Sigma_\delta^{(2)}$. We write $M_\delta = M_\delta^{(1)} \sqcup M_\delta^{(2)}$, where $M_\delta^{(j)} = S\Sigma_\delta^{(j)}$ for $j = 1, 2$, and $\partial = \partial^{(1)} \sqcup \partial^{(2)}$ with $\partial^{(j)} \subset M_\delta^{(j)}$. Note that, if $\widetilde{\mathcal{S}}_\pm^{(j)}(s)$ denotes the restriction of $\widetilde{\mathcal{S}}_\pm(s)$ to $\partial^{(j)}$, we have

$$\widetilde{\mathcal{S}}_\pm^{(1)}(s) \colon \Omega^\bullet(\partial^{(1)}) \to \mathcal{D}'^\bullet(\partial^{(2)}) \quad \text{and} \quad \widetilde{\mathcal{S}}_\pm^{(2)}(s) \colon \Omega^\bullet(\partial^{(2)}) \to \mathcal{D}'^\bullet(\partial^{(1)}).$$

As in Section 5.1,

$$\chi \widetilde{\mathcal{S}}_\pm^{(j)}(s) \chi = \chi Y_\pm^{(j)}(s) \chi + \frac{\chi \Pi_{\pm, \partial}^{(j)} \chi}{s - h_j} \quad \text{as } s \to h_j,$$

with $\mathrm{rank}(\Pi_{\pm, \partial}^{(j)}) = 1$. Here $Y_\pm^{(j)}(s)$ is holomorphic near $s = h_j$ and h_j is the topological entropy of the geodesic flow of $\Sigma_\delta^{(j)}$. As before, fix $\chi \in C_c^\infty(\partial \setminus \partial_0)$.

5.2.1 The case $h_1 \neq h_2$ We may assume $h_1 > h_2$, and we define

$$c_\pm(\chi) = \mathrm{tr}_s^\flat(\chi \widetilde{\mathcal{S}}_\pm^{(2)}(h_1) \chi^2 \Pi_{\pm,\partial}^{(1)} \chi).$$

Because $\Pi_{\pm,\partial}^{(1)}$ is of rank one, $\mathrm{tr}_s^\flat((\chi \widetilde{\mathcal{S}}_\pm^{(2)}(h_1) \chi^2 \Pi_{\pm,\partial}^{(1)} \chi)^n) = c_\pm(\chi)^n$ for any $n \geq 1$, and thus, by cyclicity of the flat trace (indeed the flat trace coincides with the real trace for operators of finite rank), as $s \to h_1$,

$$\mathrm{tr}_s^\flat((\chi \widetilde{\mathcal{S}}_\pm(s)\chi)^{2n}) = \mathrm{tr}_s^\flat((\chi \widetilde{\mathcal{S}}_\pm^{(1)}(s)\chi^2 \widetilde{\mathcal{S}}_\pm^{(2)}(s)\chi)^n + (\chi \widetilde{\mathcal{S}}_\pm^{(2)}(s)\chi^2 \widetilde{\mathcal{S}}_\pm^{(1)}(s)\chi)^n)$$
$$= \frac{2c_\pm(\chi)^n}{(s-h_1)^n} + \mathcal{O}((s-h_1)^{-n+1}).$$

Now we may proceed exactly as in Section 5.1 to obtain that, if $c_\pm(\chi) > 0$,

$$N_\pm(2n, \chi, t) \sim \frac{(c_\pm(\chi)t)^n}{n!} \frac{e^{h_\star t}}{h_\star t} \quad \text{as } t \to +\infty.$$

Remark 5.3 (continuation of Remark 5.2) If $h_1 \neq h_2$ and if we assume that $c_\pm(\chi) = 0$, then the map $s \mapsto \mathrm{tr}_s^\flat((\chi \widetilde{\mathcal{S}}_\pm(s)\chi)^2)$ has no pole on the line $\{\mathrm{Re}(s) = h_\star\}$. As in Remark 5.2, this yields

$$(5\text{-}8) \qquad\qquad N_\pm(2, \chi, t) \ll \frac{\exp(h_\star t)}{t} \quad \text{as } t \to \infty.$$

Again, the bound given in Proposition 4.5 is incompatible with (5-8) — in fact, even a weaker bound (say, a lower bound with a linear loss with respect to Theorem 2) would be incompatible with (5-8) for the case $h_1 \neq h_2$ — and this will imply that $c_\pm(\chi)$ is positive.

5.2.2 The case $h_1 = h_2 = h_\star$ In that case, by writing $c_\pm(\chi) = \mathrm{tr}_s^\flat(\chi \Pi_{\pm,\partial}^{(1)} \chi \Pi_{\pm,\partial}^{(2)})$, we have

$$\mathrm{tr}_s^\flat((\chi \widetilde{\mathcal{S}}_\pm(s)\chi)^{2n}) = \frac{2c_\pm(\chi)^n}{(s-h_\star)^{2n}} + \mathcal{O}((s-h_\star)^{-2n+1}) \quad \text{as } s \to h_\star.$$

Again, provided that $c_\pm(\chi) \neq 0$, we may proceed exactly as in Section 5.1 to obtain

$$N_\pm(2n, \chi, t) \sim 2 \frac{(c_\pm(\chi)t^2)^n}{(2n)!} \frac{e^{h_\star t}}{h_\star t}.$$

Remark 5.4 (continuation of Remark 5.3) If $h_1 = h_2$ and $c_\pm(\chi) = 0$, then the function $s \mapsto \mathrm{tr}_s^\flat((\chi \widetilde{\mathcal{S}}_\pm(s)\chi)^2)$ might have a pole at $s = h_\star$, of order at most 1. Therefore, reproducing the arguments of Section 5.1, we obtain

$$(5\text{-}9) \qquad\qquad N_\pm(2, \chi, t) = \mathcal{O}(\exp(h_\star t)) \quad \text{as } t \to \infty.$$

Note that here, assuming $c_\pm(\chi) = 0$ only wins us a factor of t for the bound on $N_\pm(2, \chi, t)$ (with respect to the asymptotics of Theorem 2), whereas in Remarks 5.2 and 5.3 we could win a bit more. This is why we need a lower bound on $N(2, L)$ which is sharp up to a sublinear loss for the case where $h_1 = h_2$ (see Proposition 4.5 and the comments below). Indeed, we will see that $N(2, t)$ can be controlled by $N_\pm(2, \chi, t)$ whenever χ has enough support; hence, Proposition 4.5 will contradict (5-9), yielding again $c_\pm(\chi) > 0$ (see Section 6.2).

6 Proof of Theorems 1 and 2

In this section we prove Theorems 1 and 2. We will apply the asymptotic growth we obtained in the last section to some appropriate sequence of functions in $C_c^\infty(\partial \setminus \partial_0)$. Let $F \in C^\infty(\mathbb{R}, [0, 1])$ be an even function such that $F \equiv 0$ on $[-1, 1]$ and $F \equiv 1$ on $]-\infty, -2] \cup [2, +\infty[$. For any small $\eta > 0$, set

$$F_\eta(t) = \sum_{k \in \mathbb{Z}} F\left(\frac{t - k\pi}{\eta}\right).$$

Then F_η is 2π–periodic and it induces a function $F_\eta \colon \mathbb{R}/2\pi\mathbb{Z} \to \mathbb{R}_{\geqslant 0}$. In the coordinates from Lemma 2.3, we define

$$\chi_\eta(z) = F_\eta(\theta) \quad \text{for } z = (\tau, 0, \theta) \in \partial.$$

Then $\chi_\eta \in C_c^\infty(\partial \setminus \partial_0)$ for any $\eta > 0$ small; the function χ_η is introduced in order to forget about trajectories passing at distance not greater than η from the "glancing set" $S\gamma_\star$.

6.1 The case $\gamma_\star$ is not separating

Recall from Section 4 that we have the a priori bounds

$$(6\text{-}1) \qquad\qquad C^{-1}\frac{e^{h_\star L}}{h_\star L} \leqslant N(1, L) \leqslant Ce^{h_\star L}$$

for L large enough. This estimate implies the following fact:[12]

$$\forall \varepsilon > 0 \ \exists L_0 > 0 \ \forall L_1 > 0 \ \exists L > L_1 \quad N(1, L - L_0) \leqslant \varepsilon N(1, L).$$

In particular, we see with the first part of (4-25) in Lemma 4.11 that, for any $\eta > 0$ small enough,

$$(6\text{-}2) \qquad\qquad \liminf_{L \to +\infty} \frac{N(1, \eta, L)}{N(1, L)} \leqslant \frac{1}{2},$$

where $N(1, \eta, L)$ is as defined in Section 4.3.

For $\eta > 0$ small and $L > 0$, neither $c_\pm(\chi_\eta)$ nor $N_\pm(n, \chi_\eta, L)$ (see Section 5.1) depend on $\pm$, since F is an even function. We denote them simply by $c(\eta)$ and $N(n, \chi_\eta, L)$, respectively. Then we claim that $c(\eta) > 0$ if $\eta > 0$ is small enough. Indeed, if $c(\eta) = 0$, then Remark 5.2 implies

$$(6\text{-}3) \qquad\qquad N(1, \chi_\eta, L) \ll \frac{\exp(h_\star L)}{h_\star L} \quad \text{as } L \to +\infty.$$

On the other hand, $N(1, L) = N(1, \chi_\eta, L) + R(\eta, L)$ with

$$R(\eta, L) = N(1, L) - N(1, \chi_\eta, L) \leqslant N(1, 2\eta, L),$$

---

[12]If it does not hold, then there is an $\varepsilon > 0$ such that, for any $L_0 > 0$, there is an L_1 such that, for any $n \geqslant 0$, it holds that $\varepsilon < N(1, L_1 + nL_0)/N(1, L_1 + (n+1)L_0)$, which gives $N(1, L_1 + (n+1)L_0)\varepsilon^n < N(1, L_1)$ for each n. Now, if L_0 is large enough, we see that (6-1) cannot hold, by making $n \to \infty$.

and thus, if η is small enough, (6-2) gives

$$\limsup_{L \to +\infty} \frac{N(1, \chi_\eta, L)}{N(1, L)} \geq \frac{1}{2}.$$

Since $C^{-1} \exp(h_\star L)/L \leq N(1, L)$ for large L, (6-3) cannot hold, and thus $c(\eta) > 0$.

In particular, we can apply Lemma 5.1 to get $\lim_L N(n, \chi_\eta, L)(n!/(c(\eta)L)^n)(h_\star L/e^{h_\star L}) = 1$. As $N(n, L) \geq N(n, \chi_\eta, L)$, for L large enough,

$$C^{-1} \frac{L^n}{n!} \frac{e^{h_\star L}}{h_\star L} \leq N(n, L) \leq C \frac{L^n}{n!} \frac{e^{h_\star L}}{h_\star L}$$

(the upper bound comes from Section 4.1.2). Let $\varepsilon > 0$. Then the above estimate combined with the second part of (4-25) in Lemma 4.11 implies that, for $\eta > 0$ small enough,

$$\limsup_L R(n, \eta, L) \frac{n!}{L^n} \frac{h_\star L}{e^{h_\star L}} < \varepsilon,$$

where $R(n, \eta, L) = N(n, L) - N(n, \chi_\eta, L)$. Writing $N(n, \chi_\eta, L) \leq N(n, L) \leq N(n, \chi_\eta, L) + R(n, \eta, L)$, we obtain

$$c(\eta)^n \leq \liminf_L N(n, L) \frac{n!}{L^n} \frac{h_\star L}{e^{h_\star L}} \leq \limsup_L N(n, L) \frac{n!}{L^n} \frac{h_\star L}{e^{h_\star L}} \leq c(\eta)^n + \varepsilon$$

for any η small enough (depending on ε!). As $\varepsilon > 0$ is arbitrary, we finally get

$$N(n, L) \sim \frac{(c_\star L)^n}{n!} \frac{e^{h_\star L}}{h_\star L} \quad \text{as } L \to +\infty,$$

where $c_\star = \lim_{\eta \to 0} c(\eta) < +\infty$ (the limit exists as $\eta \mapsto c(\eta)$ is nonincreasing and bounded by above by (6-1)).

6.2 The case $\gamma_\star$ is separating

6.2.1 The case $h_1 \neq h_2$
In this case, recall from Section 4 that we have the bound

$$\frac{C^{-1} e^{h_\star L}}{\log(L)^2} \leq N(2, L) \leq C e^{h_\star L}$$

for L large enough. In particular, using (4-26) in Lemma 4.11 and Remark 5.3, we may proceed exactly as in Section 6.1 to obtain

$$N(2n, L) \sim \frac{(c_\star L)^n}{n!} \frac{e^{h_\star L}}{h_\star L} \quad \text{as } L \to +\infty,$$

where $c_\star = \lim_{\eta \to 0} c_\pm(\chi_\eta)$.

6.2.2 The case $h_1 = h_2 = h_\star$
In this case, recall from Section 4 that we have the bound

$$\frac{C^{-1} L e^{h_\star L}}{\log(L)^4} \leq N(2, L) \leq C L e^{h_\star L}$$

for L large enough. In particular, using Lemma 4.11 and Remark 5.4, we may proceed exactly as in Section 6.1 to obtain

$$N(2n, L) \sim 2 \frac{(c_\star L)^n}{(2n)!} \frac{e^{h_\star L}}{h_\star L} \quad \text{as } L \to +\infty,$$

where $c_\star = \lim_{\eta \to 0} c_\pm(\chi_\eta)$.

7 A Bowen–Margulis type measure

7.1 Description of the constant $c_\star$

In this subsection we describe the constant $c_\star$ in terms of Pollicott–Ruelle resonant states of the open system (M_δ, φ_t), assuming for simplicity that $\gamma_\star$ is not separating. By Section 2.6, since $\Pi_{\pm,\delta}(h_\star)$ is of rank one (see Section 5.1), we may write

$$\Pi_{\pm,\delta}(h_\star)|_{\Omega^1(M_\delta)} = u_\pm \otimes (\alpha \wedge s_\mp) \quad \text{for } u_\pm \in \mathcal{D}'^1_{E^*_{\pm,\delta}}(M_\delta) \text{ and } s_\mp \in \mathcal{D}'^1_{E^*_{\mp,\delta}}(M_\delta),$$

with $\operatorname{supp}(u_\pm, s_\pm) \subset \Gamma_{\pm,\delta}$ and $u_\pm, s_\mp \in \ker(\iota_X)$. Using the Guillemin trace formula [19] and the Ruelle zeta function $\zeta_{\Sigma_\star}$, we see that the Bowen–Margulis measure μ_0 (see [9]) of the open system (M_δ, φ_t), which is given by Bowen's formula

$$\mu_0(f) = \lim_{L \to +\infty} \sum_{\substack{\gamma \in \mathcal{P}_\delta \\ \ell(\gamma) \leq L}} \frac{1}{\ell(\gamma)} \int_0^{\ell(\gamma)} f(\gamma(\tau), \dot{\gamma}(\tau)) \, d\tau \quad \text{for } f \in C_c^\infty(M_\delta),$$

coincides with the distribution $f \mapsto \operatorname{tr}^\flat_s(f \Pi_{\pm,\delta}(h)) = \int_{M_\delta} f u_\pm \wedge \alpha \wedge s_\mp$. Note that $\operatorname{supp}(u_\pm \wedge \alpha \wedge s_\mp) \subset K_\star$, where $K_\star \subset S\Sigma_\star$ is the trapped set. On the other hand, by definition of $\Pi_{\pm,\partial}$,

$$c_\star = \lim_{\eta \to 0} \operatorname{tr}^\flat_s(\chi_\eta \Pi_{\pm,\partial}) = - \lim_{\eta \to 0} \int_\partial \chi_\eta \psi^* \iota^* u_\pm \wedge \iota^* s_\mp.$$

7.2 A Bowen–Margulis type measure

In what follows we set $S_{\gamma_\star}\Sigma = \{(x, v) \in S\Sigma : x \in \gamma_\star\}$ and, for any primitive geodesic $\gamma : \mathbb{R}/\ell(\gamma)\mathbb{Z} \to \Sigma$,

$$I_\star(\gamma) = \{z \in S_{\gamma_\star}\Sigma : z = (\gamma(\tau), \dot{\gamma}(\tau)) \text{ for some } \tau\}.$$

For any $n \geq 1$, we define the set $\Gamma_n \subset S_{\gamma_\star}\Sigma$ by

$$\complement\Gamma_n = \{z \in S_{\gamma_\star}\Sigma : (\widetilde{S}_\pm)^k(z) \text{ is well defined for } k = 1, \dots, n\}.$$

Also, we set $\ell_n(z) = \max(\ell_{+,n}(z), \ell_{-,n}(z))$, where

$$\ell_{\pm,n}(z) = \ell_\pm(z) + \ell_\pm(\widetilde{S}_\pm(z)) + \cdots + \ell_\pm(\widetilde{S}_\pm^{n-1}(z)) \quad \text{for } z \in \complement\Gamma_n,$$

and $\ell_\pm(z) = \inf\{t > 0 : \varphi_{\pm t}(z) \in S_{\gamma_\star}\Sigma\}$.

We will now prove Theorem 3, which says that, for any $f \in C^\infty(S_{\gamma_\star}\Sigma)$, the limit

$$\mu_n(f) = \lim_{L \to +\infty} \frac{1}{N(n, L)} \sum_{\gamma \in \mathcal{P}_n} \frac{1}{n} \sum_{z \in I_\star(\gamma)} f(z) \tag{7-1}$$

exists and defines a probability measure μ_n on $S_{\gamma_\star}\Sigma$ supported in Γ_n. We will also prove that, in the nonseparating case,

$$(7\text{-}2) \qquad \mu_n(f) = c_\star^{-n} \lim_{\eta\to 0} \mathrm{tr}_\mathrm{s}^\flat(f(\chi_\eta \Pi_{\pm,\partial}\chi_\eta)^n),$$

where $c_\star > 0$ is the constant appearing in [Theorem 1](#). Note that here we identify f with its lift $p_\star^* f$ (which is a function on ∂), so that the above formula makes sense (recall that $p_\star \colon S\Sigma_\star \to S\Sigma$ is the natural projection which identifies both components of $\partial S\Sigma_\star = \partial$). Of course, a similar formula holds in the nonseparating case, but we omit it here.

Proof of [Theorem 3](#)   Let $f \in C^\infty(S_{\gamma_\star}\Sigma)$ be a nonnegative function. Then, reproducing the arguments in the proof of [Proposition 3.7](#), for $\mathrm{Re}(s)$ big enough,

$$\mathrm{tr}_\mathrm{s}^\flat\big(f(\chi_\eta \widetilde{\mathcal{S}}_\pm(s)\chi_\eta)^n\big) = \sum_{i(\gamma,\gamma_\star)=n} \bigg(\sum_{z\in I_\star(\gamma)} f(z)\bigg) e^{-s\ell(\gamma)} I_\star(\gamma,\chi_\eta),$$

where χ_η is as defined in [Section 6](#) and $I_\star(\gamma,\chi_\eta) = I_{\star,\pm}(\gamma,\chi_\eta)$ (see [Section 5](#); this does not depend on $\pm$, as the function F used to construct χ_η is even). Now, as f is nonnegative, we may proceed exactly as in [Section 5](#), replacing $g_{n,\chi}(t)$ by

$$g_{n,\chi_\eta,f}(t) = \sum_{\substack{\gamma\in\mathcal{P} \\ i(\gamma,\gamma_\star)=n}} \bigg(\sum_{z\in I_\star(\gamma)} f(z)\bigg) \sum_{\substack{k\geqslant 1 \\ k\ell(\gamma)\leqslant t}} I_\star(\gamma,\chi_\eta) \quad \text{for } t\geqslant 0,$$

to obtain that

$$(7\text{-}3) \qquad \lim_{L\to\infty} \frac{n!}{L^n}\frac{h_\star L}{e^{h_\star L}} \sum_{\substack{\gamma\in\mathcal{P} \\ i(\gamma_\star,\gamma)=n \\ \ell(\gamma)\leqslant L}} \bigg(\sum_{z\in I_\star(\gamma)} f(z)\bigg) I_\star(\gamma,\chi_\eta) = \mathrm{Res}_{s=h_\star} \mathrm{tr}_\mathrm{s}^\flat\big(f(\chi_\eta \widetilde{\mathcal{S}}_\pm(s)\chi_\eta)^n\big).$$

We denote by $v_{n,\eta}(f)$ the left-hand side of [(7-3)](#). Then $\eta \mapsto v_{n,\eta}(f)$ is a nonnegative and nonincreasing function which is bounded by above by $n c_\star^n \|f\|_\infty$ by [Theorem 1](#). In particular, the formula

$$\mu_n(f) = \lim_{\eta\to 0} \frac{1}{n c_\star^n} v_{n,\eta}(f) \quad \text{for } f \in C^\infty(S_{\gamma_\star}\Sigma, \mathbb{R}_{\geqslant 0})$$

defines a measure μ_n on $S_{\gamma_\star}\Sigma$ whose total mass is not greater than 1. In fact, its total mass is equal to 1, since, by definition of $c_\star$,

$$\mu_n(1) = \lim_{\eta\to 0} \frac{n c_\pm(\chi_\eta)^n}{n c_\star^n} = 1.$$

Let $\varepsilon > 0$. Then, for each $f \in C^\infty(S_{\gamma_\star}\Sigma, \mathbb{R}_{\geqslant 0})$, one has, by [Lemma 4.11](#),

$$\sum_{\substack{\gamma\in\mathcal{P} \\ i(\gamma_\star,\gamma)=n \\ \ell(\gamma)\leqslant L}} \bigg(\sum_{z\in I_\star(\gamma)} f(z)\bigg)(1 - I_\star(\gamma,\chi_\eta)) \leqslant n N(n,\eta,L)\|f\|_\infty \leqslant \varepsilon n N(n,L)\|f\|_\infty$$

for large L whenever η is small enough. In particular, setting

$$\mu_n^+(f) = \limsup_L \frac{A_f(n, L)}{nN(n, L)} \quad \text{and} \quad \mu_n^-(f) = \liminf_L \frac{A_f(n, L)}{nN(n, L)},$$

where

$$A_f(n, L) = \sum_{\substack{\gamma \in \mathcal{P} \\ i(\gamma_\star, \gamma) = n \\ \ell(\gamma) \leqslant L}} \left(\sum_{z \in I_\star(\gamma)} f(z) \right),$$

we see that, for each $\varepsilon > 0$ and η small depending on ε,

$$|\mu_n^\pm(f) - \nu_{n,\eta}(f)| \leqslant \varepsilon \|f\|_\infty.$$

Indeed, setting

$$A_f(n, \eta, L) = \sum_{\substack{\gamma \in \mathcal{P} \\ i(\gamma_\star, \gamma) = n \\ \ell(\gamma) \leqslant L}} \left(\sum_{z \in I_\star(\gamma)} f(z) \right) I_\star(\gamma, \chi_\eta),$$

we have

$$\limsup_L \left| \left(\frac{1}{nN(n, L)} - \frac{n! L^n}{n c_\star^n e^{h_\star L}} \right) A_f(n, \eta, L) \right| = 0$$

by Theorem 1, since $A_f(n, \eta, L) \leqslant nN(n, L)$. Now we may let $\eta \to 0$ to get $|\mu_n^\pm(f) - \mu_n(f)| \leqslant \varepsilon \|f\|_\infty$; since ε is arbitrary, this yields $\mu_n^\pm(f) = \mu_n(f)$. This implies that the limit (7-1) exists, and moreover (7-2) holds by (7-3) (provided that $\gamma_\star$ is not separating).

Next, take a general $f \in C^\infty(S_{\gamma_\star}\Sigma)$, which we no longer assume to be nonnegative. Choose some smooth functions $f_{\delta,\pm}$, $\delta \in {]}0, 1{[}$ with the property that $\|f - (f_{\delta,+} + f_{\delta,-})\|_\infty \leqslant \delta$ and $\pm f_{\delta,\pm} \geqslant 0$, and write $f_\delta = f_{\delta_+} + f_{\delta_-}$. By nonnegativeness of $\pm f_{\delta,\pm}$, the arguments above imply that $A_{f_\delta}(n, L)/(nN(n, L)) \to \mu_n(f_\delta)$ as $L \to \infty$. On the other hand, $|A_f(n, L) - A_{f_\delta}(n, L)| \leqslant A_{|f - f_\delta|}(n, L) \leqslant \delta nN(n, L)$. Letting $L \to \infty$, this yields

$$\mu_n(f_\delta) - \delta \leqslant \liminf_L \frac{A_f(n, L)}{nN(n, L)} \leqslant \limsup_L \frac{A_f(n, L)}{nN(n, L)} \leqslant \mu_n(f_\delta) + \delta.$$

Since $\mu_n(f_\delta) \to \mu_n(f)$ as $\delta \to 0$, (7-1) and (7-2) are valid for f.

Finally, if $f \in C_c^\infty(S_{\gamma_\star}\Sigma \setminus \Gamma_n)$ then there is $L > 0$ such that

$$\ell_n(z) \leqslant L \quad \text{for } z \in \operatorname{supp}(f).$$

In particular, for any $\gamma \in \mathcal{P}$ such that $i(\gamma, \gamma_\star) = n$ and $\ell(\gamma) > L$, we have $f(z) = 0$ for any $z \in I_\star(\gamma)$. This shows that $\mu_n(f) = 0$, and the support condition for μ_n follows. $\qquad\square$

8 A large deviation result

The goal of this section, which is independent of the rest of the paper, is to prove the following result, which is a consequence of a classical large deviation result by Kifer [25]:

Proposition 8.1 *There exists $I_\star > 0$ such that the following holds. For any $\varepsilon > 0$, there are $C, \delta > 0$ such that, for large L,*

$$(8\text{-}1) \qquad \frac{1}{N(L)} \#\left\{ \gamma \in \mathcal{P} : \ell(\gamma) \leqslant L \text{ and } \left| \frac{i(\gamma, \gamma_\star)}{\ell(\gamma)} - I_\star \right| \geqslant \varepsilon \right\} \leqslant C \exp(-\delta L).$$

In fact, $I_\star = 4i(\bar{m}, \delta_{\gamma_\star})$, where i is Bonahon's intersection form [6], $\delta_{\gamma_\star}$ is the Dirac measure on $\gamma_\star$ and $\bar{m}$ is the renormalized Bowen–Margulis measure on M (here we see the intersection form as a function on the space of φ–invariant measures on $S\Sigma$, as described below). Lalley [28] showed a similar result for self-intersection numbers; see also [41] for self-intersection numbers with prescribed angles.

8.1 Bonahon's intersection form

Let $\mathcal{M}_\varphi(S\Sigma)$ be the set of finite positive measures on $S\Sigma$ invariant by the geodesic flow, endowed with the vague topology. For any closed geodesic γ, we denote by $\delta_\gamma \in \mathcal{M}_\varphi(S\Sigma)$ the Lebesgue measure of γ parametrized by arc length (thus of total mass $\ell(\gamma)$). Let $\mu \in \mathcal{M}_\varphi(S\Sigma)$ be the Liouville measure, that is, the measure associated to the volume form $\frac{1}{2}\alpha \wedge d\alpha$.

Proposition 8.2 (Bonahon [7]; see also Otal [34]) *There exists a continuous function*

$$i : \mathcal{M}_\varphi(S\Sigma) \times \mathcal{M}_\varphi(S\Sigma) \to \mathbb{R}_+$$

which is additive and positively homogeneous with respect to each variable and such that $i(\mu, \mu) = 2\pi \operatorname{vol}(\Sigma)$ and

$$i(\delta_\gamma, \delta_{\gamma'}) = i(\gamma, \gamma') \quad \text{and} \quad i(\mu, \delta_\gamma) = 2\ell(\gamma),$$

for any closed geodesics γ and γ'.

Remark 8.3 (i) Actually, Bonahon's intersection form is a pairing on the space of *geodesic currents*. This space is naturally identified with the space of φ–invariant measures on $S\Sigma$ which are also invariant by the flip $R : (x, v) \mapsto (x, -v)$. By $i(\nu, \nu')$ for general $\nu, \nu' \in \mathcal{M}_\varphi(S\Sigma)$ we simply mean $i(\Phi(\nu), \Phi(\nu'))$ where $\Phi : \nu \mapsto \nu + R^*\nu$ (note that $\varphi_t R = R\varphi_{-t}$ for $t \in \mathbb{R}$).

(ii) The formulae for $i(\mu, \mu)$ and $i(\mu, \delta_\gamma)$ differ from [7]; this is due to our convention, since here the Liouville measure μ corresponds to twice the Liouville current considered in [7].

8.2 Large deviations

For any $\nu \in \mathcal{M}_\varphi(S\Sigma)$ we denote by $h(\nu)$ the measure-theoretical entropy of φ with respect to ν. Then we have the following result:

Proposition 8.4 (Kifer [25]) *Let $F \subset \mathcal{M}_\varphi^1(S\Sigma)$ be a closed set, where $\mathcal{M}_\varphi^1(S\Sigma)$ is the set of φ–invariant probability measures on $S\Sigma$. Then*

$$\limsup_L \frac{1}{L} \log \frac{1}{N(L)} \#\left\{ \gamma \in \mathcal{P} : \ell(\gamma) \leqslant L \text{ and } \frac{\delta_\gamma}{\ell(\gamma)} \in F \right\} \leqslant \sup_{\nu \in F} h(\nu) - h,$$

where h is the entropy of the geodesic flow.

Proof of Proposition 8.1 We denote by $\bar{m} \in \mathcal{M}^1_\varphi(S\Sigma)$ the unique probability measure of maximal entropy, that is,

$$\bar{m} = \lim_{L \to +\infty} \sum_{\substack{\gamma \in \mathcal{P} \\ \ell(\gamma) \leqslant L}} \frac{\delta_\gamma}{\ell(\gamma)},$$

where the convergence holds in the weak sense. Let $\varepsilon > 0$. Define

$$F_\varepsilon = \{\nu \in \mathcal{M}^1_\varphi(S\Sigma) : |i(\nu, \delta_{\gamma_\star}) - i(\bar{m}, \delta_{\gamma_\star})| \geqslant \varepsilon\}.$$

Then F_ε is closed in $\mathcal{M}^1_\varphi(S\Sigma)$, and thus compact by the Banach–Alaoglu theorem, and $\bar{m} \in \complement F_\varepsilon$ so that $\delta = h - \sup_{\nu \in F_\varepsilon} h(\nu) > 0$. In particular, for large L,

$$\frac{1}{N(L)} \#\left\{\gamma \in \mathcal{P} : \frac{\delta_\gamma}{\ell(\gamma)} \in F_\varepsilon\right\} \leqslant C \exp(-\delta' L)$$

for some $0 < \delta' < \delta$ and $C > 0$. Now, by Proposition 8.2, $\delta_\gamma/\ell(\gamma) \in F_\varepsilon$ gives $|i(\gamma, \gamma_\star)/\ell(\gamma) - i(\bar{m}, \delta_{\gamma_\star})| \geqslant \varepsilon$. Let $I_\star = i(\bar{m}, \delta_{\gamma_\star})$. Then it is a well-known fact that $\bar{m}$ has full support in $S\Sigma$, which implies $I_\star > 0$ by definition of $i(\bar{m}, \delta_{\gamma_\star})$; see [34]. $\qquad\square$

Remark 8.5   (i) It is not hard to see that Proposition 8.1 implies

$$\frac{1}{N(L)} \sum_{\ell(\gamma) \leqslant L} i(\gamma, \gamma_\star) \sim I_\star L$$

as $L \to +\infty$. Thus we recover [39, Theorem 4].

(ii) If (Σ, g) is hyperbolic, then $\bar{m}$ is the renormalized Liouville measure and, with Proposition 8.2, we find

$$I_\star = \frac{\ell(\gamma_\star)}{2\pi^2(g-1)}.$$

(iii) If $\varepsilon < I_\star$ then every closed geodesic γ which does not intersect $\gamma_\star$ satisfies $\delta_\gamma/\ell(\gamma) \in F_\varepsilon$. In particular, the right-hand side of (8-1) is bounded from below by $C \exp((h_\star - h)L)$, where we used that $N(0, L) \sim \exp(h_\star L)/h_\star L$ and $N(L) \sim \exp(hL)/hL$ as $L \to \infty$.

Appendix A   Closed geodesics minimize intersection numbers

In this section we prove Lemma 2.1. We proceed by contradiction and assume that $i(\gamma_1, \gamma_2) < |\gamma_1 \cap \gamma_2|$. As γ_1 and γ_2 are not powers of each other, the images of γ_1 and γ_2 intersect transversally (otherwise their images would coincide by uniqueness of the geodesic equation). Since $i(\gamma_1, \gamma_2) < |\gamma_1 \cap \gamma_2|$, we may find loops $\alpha_j : \mathbb{R}/\mathbb{Z} \to \Sigma$ for $j = 1, 2$ with $\alpha_j \sim \gamma_j$ and $|\alpha_1 \cap \alpha_2| < |\gamma_1 \cap \gamma_2|$, and we may moreover assume that α_1 and α_2 intersect transversally. Let $H_j : [0, 1] \times \mathbb{R}/\mathbb{Z} \to \Sigma$ for $j = 1, 2$ be smooth homotopies between γ_j and α_j, and define $H : [0, 1] \times \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z} \to \Sigma \times \Sigma$ by setting

$$H(s, \tau_1, \tau_2) = (H_1(s, \tau_1), H_2(s, \tau_2)) \quad \text{for } (s, \tau_1, \tau_2) \in [0, 1] \times \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}.$$

Let $\Delta(\Sigma) = \{(x,x) : x \in \Sigma\}$ be the diagonal in Σ. Then $H(0,\cdot)$ and $H(1,\cdot)$ are transverse to $\Delta(\Sigma)$, in the sense that, for every $k = 0, 1$ and $(\tau_1, \tau_2) \in \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$ with $H(k, \tau_1, \tau_2) \in \Delta(\Sigma)$,

$$\mathrm{d}H(k, \tau_1, \tau_2)T_{(k,\tau_1,\tau_2)}(\mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}) + T_{H(k,\tau_1,\tau_2)}\Delta(\Sigma) = T_{H(k,\tau_1,\tau_2)}(\Sigma \times \Sigma).$$

In particular, by [20, Corollary page 73] we may assume that H is globally transverse to $\Delta(\Sigma)$, so that $H^{-1}(\Delta(\Sigma))$ is a smooth 1–dimensional submanifold of $[0,1] \times (\mathbb{R}/\mathbb{Z})^2$. Now

$$|\gamma_1 \cap \gamma_2| = |H^{-1}(\Delta(\Sigma)) \cap (\{0\} \times (\mathbb{R}/\mathbb{Z})^2)| \quad \text{and} \quad |\alpha_1 \cap \alpha_2| = |H^{-1}(\Delta(\Sigma)) \cap (\{1\} \times (\mathbb{R}/\mathbb{Z})^2)|.$$

Since $|\gamma_1 \cap \gamma_2| > |\alpha_1 \cap \alpha_2|$ and because $H^{-1}(\Delta(\Sigma))$ is smooth, we may find a smooth path $c : [0,1] \to [0,1] \times (\mathbb{R}/\mathbb{Z})^2$ such that $c(0) \neq c(1)$ and

$$\mathrm{Im}(c) \subset H^{-1}(\Delta(\Sigma)) \quad \text{and} \quad c(0), c(1) \in \{0\} \times (\mathbb{R}/\mathbb{Z})^2.$$

Write $c = (S, T_1, T_2)$ for some smooth functions $S : [0,1] \to [0,1]$ and $T_j : [0,1] \to \mathbb{R}/\mathbb{Z}$, and for $u \in [0,1]$ define the path $c_u = (uS, T_1, T_2) : [0,1] \to [0,1] \times (\mathbb{R}/\mathbb{Z})^2$. Let $x_k = H(c(k)) \in \Sigma$ for $k = 0, 1$. Then define the paths

$$\beta_{j,u} = \pi_j \circ H \circ c_u : [0,1] \to \Sigma \quad \text{for } j = 1, 2 \text{ and } u \in [0,1],$$

where $\pi_1, \pi_2 : \Sigma \times \Sigma \to \Sigma$ are the projections over the first and second factor, respectively. As $c_1 = c$ and $\mathrm{Im}(c) \subset H^{-1}(\Delta(\Sigma))$, we have $\beta_{1,1} = \beta_{2,1}$. In particular, the paths $\beta_{1,0}$ and $\beta_{2,0}$ are homotopic within the space of curves linking x_0 and x_1, since for each u, one has $\beta_{j,u}(k) = x_k$ for $j = 1, 2$ and $k = 0, 1$. Moreover, the paths $\beta_{1,0}$ and $\beta_{2,0}$ are subpaths of γ_1 and γ_2, respectively, and are in particular geodesic paths. Let $\widetilde{\Sigma}$ be a universal cover of Σ and take $\tilde{x}_0 \in \widetilde{\Sigma}$ a lift of x_0. For $j = 1, 2$, let $\tilde{\beta}_j : [0,1] \to \widetilde{\Sigma}$ be the unique lift of $\beta_{j,0}$ starting at $\tilde{x}_0$. Then $\tilde{\beta}_1(1) = \tilde{\beta}_2(1)$ since the paths $\beta_{j,0}$ for $j = 1, 2$ are homotopic in Σ via a homotopy preserving endpoints. In particular, we have found two distinct geodesic segments of $\widetilde{\Sigma}$ joining $\tilde{x}_0$ and $\tilde{\beta}_0(1)$ (the image of the paths $\tilde{\beta}_{j,0}$ for $j = 1, 2$ cannot coincide since $c(0) \neq c(1)$ and the intersection $\gamma_1 \cap \gamma_2$ is transverse). Thus the exponential map $\exp_{\tilde{x}_0} : T_{\tilde{x}_0}\widetilde{\Sigma} \to \widetilde{\Sigma}$ at $\tilde{x}_0$ is not a diffeomorphism, and $\widetilde{\Sigma}$ cannot be negatively curved by virtue of the Cartan–Hadamard theorem (see for example [29, Theorem 11.5]). This completes the proof.

Appendix B   An elementary fact about pullbacks of distributions

Lemma B.1   *Let $K \in \mathcal{D}'(\mathbb{R}^d \times \mathbb{R}^d)$ be a compactly supported distribution. We assume that $\mathrm{WF}(K) \subset \Gamma$, where $\Gamma \subset T^*(\mathbb{R}^d \times \mathbb{R}^d)$ is a closed conical subset such that*

$$\Gamma \cap N^*\Delta = \varnothing, \quad \text{where } N^*\Delta = \{(x, \xi, x, -\xi) : (x, \xi) \in T^*\mathbb{R}^d\}.$$

In particular, the pullback i^*K, where $i : x \mapsto (x, x)$, is well defined. Then, for $N \in \mathbb{N}_{\geqslant 1}$ large enough, the following holds. Let $u \in C_c^N(\mathbb{R}^d)$ and assume that the pullback $i^*(\pi_1^* u K)$ is well defined, where $\pi_1 : (x, x) \mapsto x$ is the projection on the first factor. Then

$$i^*(\pi_1^* u \cdot K) = u \cdot i^*K.$$

Proof Let $K_\varepsilon \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d)$, $\varepsilon \in {]0, 1]}$, be a sequence of distributions supported in a fixed compact set such that $K_\varepsilon \to K$ in $\mathcal{D}'_\Gamma(\mathbb{R}^d \times \mathbb{R}^d)$. Let $\Gamma' \subset T^*(\mathbb{R}^d \times \mathbb{R}^d)$ be an open conical subset containing $N^*\Delta$. As K_ε is compactly supported, we may assume that $|t-q| > \delta_0$ for any $(t, q) \in \Gamma \times \Gamma'$ such that $|t| = |q| = 1$ for some $\delta_0 > 0$. By definition of the convergence in $\mathcal{D}'_\Gamma(\mathbb{R}^d \times \mathbb{R}^d)$ (see [23, Definition 8.2.2]), for every N there is $C_N > 0$ such that, for any $\varepsilon > 0$ small enough,

$$\text{(B-1)} \qquad |\widehat{K}_\varepsilon(q)| \leq C_N \langle q \rangle^{-N} \quad \text{for } q \in \Gamma'.$$

Let $\Gamma'' \subset \Gamma'$ be another open conical subset containing $N^*\Delta$, and let $\delta > 0$ be such that, for any $q \in \Gamma''$ and $t \in \mathbb{R}^{2d}$,

$$\text{(B-2)} \qquad |t - q| < \delta|q| \implies t \in \Gamma'.$$

Then, for any $q \in \Gamma''$,

$$
\begin{aligned}
(2\pi)^{2d} |\widehat{K_\varepsilon \pi_1^* u}(q)| &\leq \int_{\mathbb{R}_t^{2d}} |\widehat{K}_\varepsilon(t)| \cdot |\widehat{\pi_1^* u}(q-t)| \, dt \\
&\leq \int_{|t-q|<\delta|q|} |\widehat{K}_\varepsilon(t)| \cdot |\widehat{\pi_1^* u}(q-t)| \, dt + \int_{|t-q|\geq\delta|q|} |\widehat{K}_\varepsilon(t)| \cdot |\widehat{\pi_1^* u}(q-t)| \, dt.
\end{aligned}
$$

Let $N_1, N_2 \in \mathbb{N}_{\geq 1}$ and $\langle t \rangle = \sqrt{1 + |t|^2}$. Then, using (B-1), (B-2) and Peetre's inequality, and assuming that $u \in C_c^{N_2}(\mathbb{R}^d)$ with $N_2 \geq 2d + 1$,

$$
\begin{aligned}
\int_{|t-q|<\delta|t|} |\widehat{K}_\varepsilon(t)| \cdot |\widehat{\pi_1^* u}(q-t)| \, dt &\leq C_{N_1, N_2} \int_{|t-q|<\delta|q|} \langle t \rangle^{-N_1} \langle q-t \rangle^{-N_2} \, dt \\
&\leq C'_{N_1, N_2} \langle q \rangle^{-N_1 + N_2} \int_{\mathbb{R}^d} \langle t \rangle^{-N_2} \, dt.
\end{aligned}
$$

On the other hand, if k is the order of K and $N_3 \in \mathbb{N}_{\geq 1}$ is such that $u \in C_c^{N_3}(\mathbb{R}^d)$, then

$$
\begin{aligned}
\int_{|t-q|\geq\delta|q|} |\widehat{K}_\varepsilon(t)| \cdot |\widehat{\pi_1^* u}(q-t)| \, dt &\leq C_{k, N_3} \int_{|t-q|\geq\delta|q|} \langle t \rangle^k \langle q-t \rangle^{-N_3} \\
&\leq C'_{k, N_3} \langle q \rangle^{-N_3 + (k+2d+1)} \int_{\mathbb{R}^{2d}} \langle t \rangle^{-2d-1} \, dt.
\end{aligned}
$$

Therefore, if $u \in C^N(\mathbb{R}^d)$ with $N = k + 2d + 1 + N'$,

$$\text{(B-3)} \qquad (2\pi)^{2d} |\widehat{K_\varepsilon \pi_1^* u}(q)| \leq C_N \langle q \rangle^{-N'} \quad \text{for } q \in \Gamma''.$$

Note that, for $\varphi \in C_c^\infty(\mathbb{R}^d)$,

$$\langle i^*(K_\varepsilon \pi_1^* u), \varphi \rangle = \int_{\mathbb{R}_x^d} \varphi(x) \int_{\mathbb{R}_\xi^d \times \mathbb{R}_\eta^d} \widehat{K_\varepsilon \pi_1^* u}(\xi, \eta) e^{ix(\xi+\eta)} \, d\xi \, d\eta \, dx.$$

Indeed, (B-3) shows that the integral in (ξ, η) converges near $N^*\Delta$ if $N' \geq 2d + 1$, and far from $N^*\Delta$ we can use the stationary phase method to get enough convergence in (ξ, η), so the above integral makes sense as an oscillatory integral and coincides with $\langle i^*(K_\varepsilon \pi_1^* u), \varphi \rangle$, since this formula is obviously true if u is smooth. Moreover, all the above estimates are uniform in ε and thus, letting $\varepsilon \to 0$, we obtain the desired result, since obviously $i^*(K_\varepsilon \pi_1^* u) = u(i^* K_\varepsilon)$ for each $\varepsilon \in {]0, 1]}$. $\qquad \square$

References

[1] **N Anantharaman**, *Precise counting results for closed orbits of Anosov flows*, Ann. Sci. École Norm. Sup. 33 (2000) 33–56 MR Zbl

[2] **D V Anosov**, *Geodesic flows on closed Riemann manifolds with negative curvature*, Proceedings of the Steklov Institute of Mathematics 90, Amer. Math. Soc., Providence, RI (1969) MR Zbl

[3] **M F Atiyah**, **R Bott**, *A Lefschetz fixed point formula for elliptic complexes, I*, Ann. of Math. 86 (1967) 374–407 MR Zbl

[4] **V Baladi**, **M F Demers**, *On the measure of maximal entropy for finite horizon Sinai billiard maps*, J. Amer. Math. Soc. 33 (2020) 381–449 MR Zbl

[5] **V Baladi**, **M F Demers**, **C Liverani**, *Exponential decay of correlations for finite horizon Sinai billiard flows*, Invent. Math. 211 (2018) 39–177 MR Zbl

[6] **F Bonahon**, *Bouts des variétés hyperboliques de dimension* 3, Ann. of Math. 124 (1986) 71–158 MR Zbl

[7] **F Bonahon**, *The geometry of Teichmüller space via geodesic currents*, Invent. Math. 92 (1988) 139–162 MR Zbl

[8] **Y Bonthonneau**, *Les résonances du Laplacien sur les variétés à pointes*, PhD thesis, Université Paris Sud (2015) Available at `http://www.theses.fr/2015PA112141.pdf`

[9] **R Bowen**, *The equidistribution of closed geodesics*, Amer. J. Math. 94 (1972) 413–423 MR Zbl

[10] **R Bowen**, *Symbolic dynamics for hyperbolic flows*, Amer. J. Math. 95 (1973) 429–460 MR Zbl

[11] **M R Bridson**, **A Haefliger**, *Metric spaces of non-positive curvature*, Grundl. Math. Wissen. 319, Springer (1999) MR Zbl

[12] **F Dal'bo**, *Remarques sur le spectre des longueurs d'une surface et comptages*, Bol. Soc. Brasil. Mat. 30 (1999) 199–221 MR Zbl

[13] **N V Dang**, **G Rivière**, *Poincaré series and linking of Legendrian knots*, preprint (2020) arXiv 2005.13235

[14] **H Delange**, *Généralisation du théorème de Ikehara*, Ann. Sci. École Norm. Sup. 71 (1954) 213–242 MR Zbl

[15] **S Dyatlov**, **C Guillarmou**, *Pollicott–Ruelle resonances for open systems*, Ann. Henri Poincaré 17 (2016) 3089–3146 MR Zbl

[16] **S Dyatlov**, **M Zworski**, *Dynamical zeta functions for Anosov flows via microlocal analysis*, Ann. Sci. Éc. Norm. Supér. 49 (2016) 543–577 MR Zbl

[17] **V Erlandsson**, **J Souto**, *Counting curves in hyperbolic surfaces*, Geom. Funct. Anal. 26 (2016) 729–777 MR Zbl

[18] **C Guillarmou**, *Lens rigidity for manifolds with hyperbolic trapped sets*, J. Amer. Math. Soc. 30 (2017) 561–599 MR Zbl

[19] **V Guillemin**, *Lectures on spectral theory of elliptic operators*, Duke Math. J. 44 (1977) 485–517 MR Zbl

[20] **V Guillemin**, **A Pollack**, *Differential topology*, Prentice-Hall, Englewood Cliffs, NJ (1974) MR Zbl

[21] **L Guillopé**, *Sur la distribution des longueurs des géodésiques fermées d'une surface compacte à bord totalement géodésique*, Duke Math. J. 53 (1986) 827–848 MR Zbl

[22] **G Higman**, **B H Neumann**, **H Neumann**, *Embedding theorems for groups*, J. London Math. Soc. 24 (1949) 247–254 MR Zbl

[23] **L Hörmander**, *The analysis of linear partial differential operators, I: Distribution theory and Fourier analysis*, 2nd edition, Grundl. Math. Wissen. 256, Springer (1990) MR Zbl

[24] **A Katsuda**, **T Sunada**, *Homology and closed geodesics in a compact Riemann surface*, Amer. J. Math. 110 (1988) 145–155 MR Zbl

[25] **Y Kifer**, *Large deviations, averaging and periodic orbits of dynamical systems*, Comm. Math. Phys. 162 (1994) 33–46 MR Zbl

[26] **S P Lalley**, *Closed geodesics in homology classes on surfaces of variable negative curvature*, Duke Math. J. 58 (1989) 795–821 MR Zbl

[27] **S P Lalley**, *Renewal theorems in symbolic dynamics, with applications to geodesic flows, non-Euclidean tessellations and their fractal limits*, Acta Math. 163 (1989) 1–55 MR Zbl

[28] **S P Lalley**, *Self-intersections of closed geodesics on a negatively curved surface: statistical regularities*, from "Convergence in ergodic theory and probability" (V Bergelson, P March, J Rosenblatt, editors), Ohio State Univ. Math. Res. Inst. Publ. 5, de Gruyter, Berlin (1996) 263–272 MR Zbl

[29] **J M Lee**, *Riemannian manifolds: an introduction to curvature*, Graduate Texts in Math. 176, Springer (1997) MR Zbl

[30] **R C Lyndon**, **P E Schupp**, *Combinatorial group theory*, Ergebnisse der Math. 89, Springer (1977) MR Zbl

[31] **G A Margulis**, *Certain applications of ergodic theory to the investigation of manifolds of negative curvature*, Funkcional. Anal. i Priložen. 3 (1969) 89–90 MR Zbl In Russian; translated in Funct. Anal. Appl. 3 (1969) 335–336

[32] **M Mirzakhani**, *Growth of the number of simple closed geodesics on hyperbolic surfaces*, Ann. of Math. 168 (2008) 97–125 MR Zbl

[33] **M Mirzakhani**, *Counting mapping class group orbits on hyperbolic surfaces*, preprint (2016) arXiv 1601.03342

[34] **J-P Otal**, *Le spectre marqué des longueurs des surfaces à courbure négative*, Ann. of Math. 131 (1990) 151–162 MR Zbl

[35] **W Parry**, **M Pollicott**, *An analogue of the prime number theorem for closed orbits of Axiom A flows*, Ann. of Math. 118 (1983) 573–591 MR Zbl

[36] **W Parry**, **M Pollicott**, *Zeta functions and the periodic orbit structure of hyperbolic dynamics*, Astérisque 187–188, Soc. Math. France, Paris (1990) MR Zbl

[37] **F Paulin**, **M Pollicott**, **B Schapira**, *Equilibrium states in negative curvature*, Astérisque 373, Soc. Math. France, Paris (2015) MR Zbl

[38] **R Phillips**, **P Sarnak**, *Geodesics in homology classes*, Duke Math. J. 55 (1987) 287–297 MR Zbl

[39] **M Pollicott**, *Asymptotic distribution of closed geodesics*, Israel J. Math. 52 (1985) 209–224 MR Zbl

[40] **M Pollicott**, *Homology and closed geodesics in a compact negatively curved surface*, Amer. J. Math. 113 (1991) 379–385 MR Zbl

[41] **M Pollicott**, **R Sharp**, *Angular self-intersections for closed geodesics on surfaces*, Proc. Amer. Math. Soc. 134 (2006) 419–426 MR Zbl

[42] **T Roblin**, *Ergodicité et équidistribution en courbure négative*, Mém. Soc. Math. Fr. 95, Soc. Math. France, Paris (2003) MR Zbl

[43] **P C Sarnak**, *Prime geodesic theorems*, PhD thesis, Stanford University (1980) MR Available at `https://www.proquest.com/docview/303065936`

[44] **Y G Sinai**, *Dynamical systems with elastic reflections: ergodic properties of dispersing billiards*, Uspehi Mat. Nauk 25 (1970) 141–192 MR Zbl In Russian; translated in Russian Math. Surveys 25 (1970) 137–189

[45] **I M Singer**, **J A Thorpe**, *Lecture notes on elementary topology and geometry*, Scott, Foresman and Co, Glenview, IL (1967) MR Zbl

Institut de Mathématiques d'Orsay, Université Paris-Saclay
Orsay, France

Current address: *Laboratoire de Mathématiques Jean Leray, Université de Nantes*
Nantes, France

`yann.chaubet@univ-nantes.fr`

On endomorphisms of the de Rham cohomology functor

SHIZHANG LI

SHUBHODIP MONDAL

We compute the moduli of endomorphisms of the de Rham and crystalline cohomology functors, viewed as a cohomology theory on smooth schemes over truncated Witt vectors. As applications of our result, we deduce Drinfeld's refinement of the classical Deligne–Illusie decomposition result for de Rham cohomology of varieties in characteristic $p > 0$ that are liftable to W_2, and prove further functorial improvements.

1 Introduction

Let A be a ring and let X be a smooth A–scheme. The algebraic de Rham cohomology is a cohomology theory designed by Grothendieck. It is defined functorially by sending X to the hypercohomology of the de Rham complex $\Omega^*_{X/A}$. The de Rham complex $\Omega^*_{X/A}$ is not just a complex, but also has the structure of a sheaf of commutative differential graded algebras. One can therefore view the output of de Rham cohomology as a commutative algebra object in the derived ∞–category $D(A)$, which we denote by $\mathrm{CAlg}(D(A))$. This way, one obtains a functor $\mathrm{dR}_{(\cdot)/A} \colon \mathrm{Alg}^{\mathrm{sm}}_A \to \mathrm{CAlg}(D(A))$, which sends any smooth A–algebra R to $\mathrm{dR}_{R/A} \in \mathrm{CAlg}(D(A))$. Our primary goal here is to study endomorphisms of this functor.

Studying properties of the de Rham cohomology theory as a functor is interesting for a number of reasons. From a technical point of view, in certain situations, showing that the de Rham cohomology functor has no nontrivial automorphisms has been used as a key tool by Bhatt, Lurie and Mathew [7] and Li and Liu [21] to prove that certain constructions are functorially isomorphic. Further, in [24] Mondal showed

that one can reconstruct the theory of crystalline cohomology as the *unique* deformation of de Rham cohomology theory viewed as a functor defined on smooth $\mathbb{F}_p$–schemes.

From a different perspective, any property enjoyed by the de Rham cohomology functor will in particular be enjoyed by de Rham cohomology of every smooth algebraic variety. For example, if the functor $\mathrm{dR}_{(\cdot)/A}$ has many endomorphisms, one potentially obtains many interesting endomorphisms of the de Rham cohomology of any smooth algebraic variety, which could be useful for making interesting geometric conclusions. The classical study and usage of the Frobenius operator on de Rham or crystalline cohomology theory is an instance of such a perspective.

Our main motivating questions, which can be seen as a "moduli" enhancement of the question of endomorphisms of the de Rham cohomology functor, are the following:

(1) Given a ring A, what is the endomorphism monoid[1] of the functor dR that sends any smooth A–algebra R to $\mathrm{dR}_{R/A} \in \mathrm{CAlg}(D(A))$?

(2) More generally, letting B be an arbitrary A–algebra, what is the endomorphism monoid of the analogous functor $R \mapsto \mathrm{dR}_{R/A} \otimes_A B \in \mathrm{CAlg}(D(B))$?

(3) Finally, consider the presheaf[2] (of monoids) on $(A\text{–Alg})^{\mathrm{op}}$ that sends an A–algebra B to the endomorphism monoid in previous question. Is it represented by a (monoid) scheme? If so, what is the representing monoid scheme?

We address the above questions when $A = W_n(k)$ for any perfect ring k, where $W_n(k)$ denotes the ring of n–truncated Witt vectors. We expect the methods to be extendable to more general base rings but we do not pursue that direction further here.

A foretaste of the main theorem

For simplicity, let us focus now on the case where $A = \mathbb{Z}/p^n$ or $\mathbb{Z}_p$ and B is an $\mathbb{F}_p$–algebra.

Theorem 1.1 (special case of Theorem 4.24, the main theorem) (1) *When $A = \mathbb{F}_p$, the endomorphism monoid of $\mathrm{dR}_{(\cdot)/A} \otimes_A B$ is $\underline{\mathbb{N}}(\mathrm{Spec}(B))$, where $\underline{\mathbb{N}}$ denotes the constant monoid scheme associated with the natural numbers.*

(2) *However, when $A = \mathbb{Z}/p^n$ for $n \geq 2$, the endomorphism monoid of $\mathrm{dR}_{(\cdot)/A} \otimes_A B$ is a semidirect product of $\underline{\mathbb{N}}(\mathrm{Spec}(B))$ with a group $W(B)^\times[F]$, the Frobenius kernel of the unit group in $W(B)$.*

Remark 1.2 (1) Roughly speaking, when $A = \mathbb{Z}/p^n$ for $n \geq 2$, Theorem 1.1 says that the endomorphism monoid of dR is very large. More precisely, Theorem 1.1 provides an action of $W^\times[F]$ on the mod p de Rham cohomology of a variety liftable to W_2. Recently, Drinfeld has also observed an action of $W^\times[F]$ on the mod p de Rham cohomology, using his (and, independently, Bhatt and Lurie's) theory of "prismatization". The main new ingredient of Theorem 1.1 is to go *beyond* this action and *classify all the*

---

[1] A priori we get a monoid object in spaces rather than an actual monoid. But in the cases of interest to us, this space is discrete; see Lemmas 3.3 and 4.2.

[2] Mathew pointed out to us that this presheaf is automatically an fpqc sheaf by flat descent.

endomorphisms. Interestingly, our proof of Theorem 1.1 does not make any use of prismatization, and only uses the stacky approach to de Rham cohomology theory in positive characteristic that already appeared in work of Drinfeld [13]. However, while the stacky approach (including the theory of prismatization) helps in constructing the endomorphisms, it does not a priori offer any strategy to prove that they are *all* the endomorphisms. To achieve this, we employ some very different additional techniques in the proof of Theorem 1.1, such as the theory of affine stacks due to Toën [29], a version of the topos-theoretic cotangent complex (see Appendix A) due to Illusie [17], and some explicit computations when necessary.

(2) The $W^\times[F]$ action resulting from Theorem 1.1 will be utilized to prove a strengthened version of the Deligne–Illusie decomposition; see Theorem 1.6. See Corollary 1.7 for an application of the full classification offered by Theorem 1.1.

(3) From the above calculation, one finds that, for $A = \mathbb{Z}/p^n$, the association $B \mapsto \mathrm{End}(\mathrm{dR}_{(\cdot)/A} \otimes_A B)$ defines a sheaf of monoids representable by a scheme denoted by $\mathrm{End}_{1,n}$. The representing monoid scheme depends on $A = \mathbb{Z}/p^n$ and stabilizes when $n \geq 2$.

The stabilization we refer to means the following: Observe that we have a natural commutative diagram

$$
\begin{array}{ccc}
\mathrm{Alg}^{\mathrm{sm}}_{\mathbb{Z}/p^n} & \xrightarrow{\ \mathrm{mod}\ p^{n-1}\ } & \mathrm{Alg}^{\mathrm{sm}}_{\mathbb{Z}/p^{n-1}} \\
\ _{\mathrm{dR}\otimes B}\searrow & & \swarrow_{\mathrm{dR}\otimes B} \\
& \mathrm{CAlg}(D(B)) &
\end{array}
$$

which induces a sequence of maps of schemes

$$\mathrm{End}_{1,1} \to \mathrm{End}_{1,2} \to \cdots \to \mathrm{End}_{1,n} \to \cdots.$$

Our theorem says the first map is a closed immersion, and all subsequent maps are isomorphisms.

Remark 1.3 The representing monoid scheme stabilizes as soon as A leaves characteristic p; this indicates that the functorial Frobenius endomorphism is solely responsible for the rigidity of de Rham cohomology theory in characteristic p.

Regarding endomorphisms of de Rham cohomology itself, we also get the following:

Theorem 1.4 (special case of Proposition 3.5) *When $A = \mathbb{Z}_p$, the endomorphism monoid of $\mathrm{dR}^\wedge_{(\cdot)/A}$ is $\mathbb{N}$, given by powers of the Frobenius.*

Here the $\mathrm{dR}^\wedge$ denotes the p–adic derived de Rham cohomology theory; see Bhatt [3]. The fact that there is no automorphism of p–adic derived de Rham cohomology theory when the base ring is p–complete and p–torsion-free was observed by Li and Liu [21, Theorem 3.14].

Remark 1.5 In both cases $A = \mathbb{F}_p$ and $\mathbb{Z}_p$ above, we only see powers of the Frobenius as endomorphisms of the (p–adic) de Rham cohomology, but this is for two different reasons: when $A = \mathbb{F}_p$ it is due to the existence of the Frobenius endomorphism on the category of A–algebras, whereas for $A = \mathbb{Z}_p$ it comes from the fact that A is p–torsion-free, so a certain huge group scheme has no nontrivial A–valued point.

In Theorem 4.24 we work in a more general setting. Namely we calculate the moduli of endomorphisms of crystalline cohomology theory, leading to sheaves $\text{End}_{m,n}$ (see Corollary 4.27 for the precise statement). The result is similar: the Frobenius endomorphisms that people "knew and loved" correspond to the monoid underlying the connected components of the whole endomorphism monoid. In fact, there is a distinguished point in each component which corresponds to a power of the Frobenius endomorphism. Furthermore, the identity component also stabilizes to a large and mysterious group scheme (see Definition 4.14), which demands further investigations (see Remark 4.30). One surprising feature is that the above group scheme is nonflat over the base in the general setting of crystalline cohomology.

Application to the Deligne–Illusie decomposition

As an application of the $W^\times[F]$–action, Drinfeld observed a refinement of the Deligne–Illusie decomposition, which was communicated to us by Bhatt (see Bhatt and Lurie [5, Example 4.7.17 and Remark 4.7.18; 6, Remark 5.16]): since $\mu_p \subset W^\times[F]$, the mod p de Rham cohomology of varieties liftable to W_2 has the structure of a μ_p–representation. It is easy to see that the $W^\times[F]$–action preserves conjugate filtration. Then one needs to show that the ith graded piece of conjugate filtration is pure of weight $i \in \mathbb{Z}/p$ as a μ_p–representation. In [5; 6], this statement is proven by establishing a relation between the $W^\times[F]$–action and the "Sen operator" defined in loc. cit. In Theorem 5.4, we use a more direct argument to check that the weight statement holds for the $W^\times[F]$–action coming from our Theorem 1.1.

Theorem 1.1, coupled with the calculation of weights from Theorem 5.4 as above, immediately implies the following improvement of results due to Achinger and Suh [1, Theorem 1.1], which in turn is a strengthening of Deligne and Illusie's result [12, corollaire 2.4]. In particular, our approach gives a proof, different from Bhatt and Lurie's, of the following result, which does not make any use of prismatization:

Theorem 1.6 (Bhatt and Lurie [6] and Drinfeld; see Corollary 5.6) *Let k be a perfect ring of characteristic $p > 0$, let X be a smooth scheme over $W_2(k)$, and let $a \le b \le a + p - 1$. Then the canonical truncation $\tau_{[a,b]}(\Omega^\bullet_{X_k/k})$ splits. Moreover, the splitting is functorial in the lift X of X_k.*

Since our calculation shows the endomorphism monoid of mod p de Rham cohomology stabilizes after W_2, philosophically it says that further liftability over W_n for $n > 2$ provides no extra knowledge on the mod p de Rham cohomology.

It is still an open problem whether there exists a smooth variety X (necessarily of dimension $\dim X > p$) over k which lifts to $W_2(k)$ for which the de Rham complex is not decomposable.[3] Using Theorem 1.1, we obtain a somewhat negative result in this direction: we show that the de Rham complex of smooth varieties over k liftable to $W_2(k)$ does not completely decompose in a *functorial* manner as a commutative algebra object in the derived category.

---

[3] A counterexample has recently been constructed by Petrov [27].

Corollary 1.7 (see Proposition 4.29) *There is no functorial splitting*

$$\mathrm{dR}_{(-\otimes_{W_2(k)}k)/k} \simeq \bigoplus_{i \in \mathbb{N}_{\geq 0}} \mathrm{Gr}_i^{\mathrm{conj}}(\mathrm{dR}_{(-\otimes_{W_2(k)}k)/k})$$

as a functor from smooth $W_2(k)$–algebras to $\mathrm{CAlg}(D(k))$.

The above statement was also observed by Mathew. His idea for a proof does not use the full calculation of endomorphism monoids as in Theorem 1.1, whereas for us it is a consequence of that calculation.

Lastly, one may wonder if the Drinfeld splitting agrees with the Deligne–Illusie splitting (which has an ∞–categorical functorial enhancement; see Kubrak and Prikhodko [20, Theorem 1.3.21 and Proposition 1.3.22]). Both splittings are obtained from the splitting of the first conjugate filtration via an averaging process; see step (a) in Deligne and Illusie's proof of [12, théorème 2.1]. To guarantee that the above two splittings are functorially the same, we show the following uniqueness:

Theorem 1.8 (see Theorem 5.10 for the precise statement) *There is a unique functorial splitting* (*as $\mathrm{Fil}_0^{\mathrm{conj}}$–modules*)

$$\mathrm{Fil}_1^{\mathrm{conj}}(\mathrm{dR}_{(-\otimes_{W_2(k)}k)/k}) = \mathrm{Fil}_0^{\mathrm{conj}}(\mathrm{dR}_{(-\otimes_{W_2(k)}k)/k}) \oplus \mathrm{Gr}_1^{\mathrm{conj}}(\mathrm{dR}_{(-\otimes_{W_2(k)}k)/k}).$$

In particular, the Deligne–Illusie splitting of Kubrak and Prikhodko [20], the Drinfeld splitting of Bhatt and Lurie [5; 6], and the splitting induced by Theorem 1.1 must all agree.

Outline of the proof of Theorem 1.1

Let us briefly outline the key ingredients in the proof of Theorem 1.1. In doing so, we will also give a rough outline of the paper. For simplicity, let us fix $A = \mathbb{Z}/p^n$, and let B be an $\mathbb{F}_p$–algebra.

(0) Theorem 1.4 is within reach of the quasisyntomic descent techniques introduced by Bhatt, Morrow and Scholze [9]; see Section 3. We also use quasisyntomic descent techniques to show that the endomorphism spaces of interest to us are actually discrete (see Lemmas 3.3 and 4.2).

(1) For Theorem 1.1(2), we need to make use of the stacky approach to de Rham or crystalline cohomology due to Drinfeld [13; 14], which can be seen as a positive-characteristic analogue of Simpson's de Rham stack; see Simpson [28]. Here we use a compressed version of the stacky approach: the functor $\mathrm{dR}_{(\cdot)/A} \otimes_A B$ is built as the *unwinding* (see Section 2.4) of an A–algebra stack over B; this stack is denoted by $\mathbb{A}_B^{1,\mathrm{dR}}$ (we often omit B to ease the notation). This unwinding construction is a variant of a construction used by Mondal [24, Section 3]. Note the amusing switch of roles played by A and B: the de Rham cohomology theory is a cohomology theory for varieties over A with coefficient ring being B, whereas the stack $\mathbb{A}_B^{1,\mathrm{dR}}$ is an A–algebra object over B.

(2) It turns out that the underlying stack $\mathbb{A}^{1,\mathrm{dR}}$ is an affine stack, in the sense of Toën [29]. Roughly speaking, for affine stacks one can pass to the "ring" of derived global sections in a lossless manner. Using this property, in Proposition 4.4 we show that $\mathrm{End}(\mathrm{dR}_{(\cdot)/A} \otimes_A B) \simeq \mathrm{End}_{A-\mathrm{Alg}-\mathrm{St}}(\mathbb{A}_B^{1,\mathrm{dR}})$. Here the latter endomorphisms are taken in the category of A–algebra stacks over B.

(3)   Using the description of $\mathbb{A}^{1,\mathrm{dR}}$ as the quotient stack $[W/pW]$, where W denotes the ring scheme of p–typical Witt vectors, in Section 4.2 we construct "enough" endomorphisms of $\mathrm{dR}_{(\cdot)/A} \otimes_A B$ and show that the endomorphism monoid is at least as big as Theorem 1.1 claims.

(4)   To finish the proof of Theorem 1.1, one needs to show that there are no endomorphisms other than the ones already constructed. To do so, we interpret an endomorphism of the algebra stack $\mathbb{A}^{1,\mathrm{dR}}$ as a deformation of an endomorphism of the sheaf of rings $\pi_0(\mathbb{A}^{1,\mathrm{dR}})$. We know that (see Proposition 4.20) $\pi_0(\mathbb{A}^{1,\mathrm{dR}}_B) = \mathbb{G}_{a,B}$ because B is an $\mathbb{F}_p$–algebra. Then we use the formalism of topos-theoretic cotangent complexes due to Illusie [17] (see Appendix A) to understand this deformation problem. This is carried out in Theorem 4.24, where we use the cotangent complex and the transitivity triangle to finish calculating the desired endomorphism monoid.

Remark 1.9   Let A and B be as above. Combining steps (2) and (3), an endomorphism of the functor $\mathrm{End}(\mathrm{dR}_{(\cdot)/A} \otimes_A B)$ is the same datum as a natural endomorphism of $W(S)/^L p$, as an animated A–algebra, for every (discrete) B–algebra S.

Acknowledgements

2   Stacky approach to de Rham cohomology

The goal of this section is to describe the stacky approach to de Rham cohomology theory due to Drinfeld [13]. Roughly, given a scheme X, Drinfeld constructed a stack X^{dR} such that $R\Gamma(X^{\mathrm{dR}}, \mathcal{O})$ recovers the de Rham cohomology $R\Gamma_{\mathrm{dR}}(X)$. This should be seen as a positive characteristic variant of the earlier construction of the de Rham stack due to Simpson [28].

For our purposes, we will need to work with a certain compressed version of this construction. Our goal is to consider a *single* stack with enough structure encoded, which can naturally "unwind" itself to construct the stack X^{dR} for *every* scheme X. To this end, we will begin by discussing quasi-ideals (see [14, Section 3.1; 24, Section 3.2]) and ring stacks, which formulates exactly the kind of extra structures on a stack one needs to work with in order to use the unwinding machine. After that, we will discuss the construction of this unwinding functor, and explain how to build a cohomology theory from a ring stack

in general. We will then discuss the particular ring stack $\mathbb{A}^{1,\mathrm{dR}}$ which gives rise to de Rham cohomology theory via this construction. For later application, the fact that the stack $\mathbb{A}^{1,\mathrm{dR}}$ is an *affine stack* in the sense of [29] will be of particular importance to us. Therefore we will record the relevant definitions in this section as well.

2.1 Quasi-ideals

Definition 2.1 (quasi-ideals) Let R be a ring and M be an R–module equipped with a map $d : M \to R$ of R–modules which satisfies $d(x) \cdot y = d(y) \cdot x$ for any pair $x, y \in M$. Such a data $d : M \to R$ satisfying the aforementioned condition will be called a *quasi-ideal* in R, or simply a quasi-ideal.

A morphism of quasi-ideals $(d_1 : M_1 \to R_1) \to (d_2 : M_2 \to R_2)$ is defined to be a pair of maps $a : M_1 \to M_2$ and $b : R_1 \to R_2$ such that the following compatibilities hold:

(1) $d_2 a = b d_1$.

(2) $a(r_1 m_1) = b(r_1) a(m_1)$.

(3) b is a ring homomorphism.

(4) a is linear.

In other words, we want a commutative diagram

$$
\begin{array}{ccc}
M_1 & \xrightarrow{\ a\ } & M_2 \\
d_1 \downarrow & & \downarrow d_2 \\
R_1 & \xrightarrow{\ b\ } & R_2
\end{array}
$$

such that b is a ring homomorphism and a is an R_1–module map $M_1 \to b_* M_2$. The category of quasi-ideals will be denoted by QID.

Construction 2.2 (quasi-ideal as a simplicial abelian group) Given a quasi-ideal $(d : M \to R)$, we obtain a map $t : T := M \times R \to R$ given by $(m, r) \mapsto r + d(m)$. There is another map $s : M \times R \to R$ given by $(m, r) \mapsto r$. There is also a degeneracy map $e : R \to M \times R$ given by $r \mapsto (0, r)$. Lastly, there is a map $c : T \times_{R,s,t} T \to T$ given by

$$(r, m) \times (r', m') \mapsto (r, m + m'),$$

where $t(r, m) = s(r', m')$ so that $(r, m) \times (r', m') \in T \times_{R,s,t} T$. Therefore we obtain a groupoid denoted by

$$M \times R \rightrightarrows R.$$

Note that the morphisms s, t, c and e are morphisms of abelian groups, so one can actually convert the above data into a 1–truncated simplicial abelian group.

In the construction below, we explain how to attach a 1–truncated simplicial ring or a ring groupoid from the data of a quasi-ideal.

Construction 2.3 (quasi-ideal as a simplicial commutative ring) Let $d : M \to R$ be a quasi-ideal. We have already defined a groupoid

$$M \times R \rightrightarrows R,$$

which can also be thought of as a 1–truncated simplicial abelian group. Next, we give a ring structure on $M \times R$. We define $(m_1, r_1) \cdot (m_2, r_2) := (r_2 m_1 + r_1 m_2 + d(m_1) m_2, r_1 r_2)$. Now, as one easily checks, the morphisms s, t, c and e in the definition of the groupoid

$$M \times R \rightrightarrows R$$

are all *ring* homomorphisms with respect to the ring structure on $M \times R$ defined above. The above data can be converted into a 1–truncated simplicial commutative ring.

Definition 2.4 (quasi-ideals in schemes) Let R be a ring scheme and M be a module scheme over R equipped with a map $d : M \to R$ of R–module schemes. This data will be called a *quasi-ideal in R* if $d(x) \cdot y = d(y) \cdot x$ for scheme-theoretic points $x, y \in M$.

A morphism between quasi-ideals in schemes is defined in a way similar to Definition 2.1.

Finally, let us give some examples of quasi-ideals that will be used later on. For more details on these examples, we refer the reader to [14, Sections 3.2–3.5] or [24, Section 2.2].

Example 2.5 Let $\mathbb{G}_a^{\sharp} \to \mathbb{G}_a$ denote the quasi-ideal obtained by taking the divided power envelope of the origin inside $\mathbb{G}_a$.

Example 2.6 Let B be any ring on which p is nilpotent. Then the functor $S \to S^{\flat} := \varprojlim_F S/p$ is representable by the affine ring scheme $\operatorname{Spec} B[x^{1/p^{\infty}}]$, which will be denoted by $\mathbb{G}_a^{\mathrm{perf}}$.

Example 2.7 Let $\mathbb{G}_a^{\mathrm{perf}, \sharp} \to \mathbb{G}_a^{\mathrm{perf}}$ denote the quasi-ideal obtained by taking the divided power envelope of the closed subscheme defined by the ideal (p, x) inside $\mathbb{G}_a^{\mathrm{perf}}$ compatibly with the existing divided powers of p.

Example 2.8 Let W denote the ring scheme of p–typical Witt vectors. By taking the kernel of the Frobenius F, one obtains a quasi-ideal $W[F] \to \mathbb{G}_a$, which is isomorphic to $\mathbb{G}_a^{\sharp} \to \mathbb{G}_a$ as a quasi-ideal in $\mathbb{G}_a$.

Example 2.9 By considering the multiplication by p map on W, one obtains a quasi-ideal $W \xrightarrow{\times p} W$.

2.2 Ring stacks

We begin by collecting some notation. If C and D denote two ∞–categories which have finite products, then the category of finite product preserving functors will be given by $\operatorname{Fun}_{\times}(C, D)$. Let Poly_A denote the category of finitely generated polynomial algebras over A.

Definition 2.10 (animated ring objects in a category) Let C be an ∞–category with products. Animated A–algebra objects in C, denoted by $\operatorname{ARings}(C)_A$, is defined to be the category $\operatorname{Fun}_{\times}(\operatorname{Poly}_A^{\mathrm{op}}, C)$.

In the case where C is the ∞–category of spaces, then the above definition with $A = \mathbb{Z}$ recovers the usual category of animated rings.

Remark 2.11 The ∞–category of animated rings has all small colimits. Given a simplicial commutative ring, one can take the colimit over the simplex category and obtain an animated ring. In particular, given a quasi-ideal, one can apply Construction 2.3 and obtain an animated ring.

Definition 2.12 (prestacks) The ∞–category of prestacks over a fixed (discrete) base ring B, denoted by PreSt_B, is defined to be the category of functors $\mathrm{Fun}(\mathrm{Alg}_B, \mathcal{S})$, where Alg_B is the category of discrete B–algebras and $\mathcal{S}$ is the ∞–category of spaces.

We note that even though we do not impose any sheafiness conditions, the examples of stacks we consider will all be (hypercomplete) fpqc sheaves of spaces.

Definition 2.13 (A–algebra prestacks over $\mathrm{Spec}(B)$) The category of A–algebra prestacks over $\mathrm{Spec}(B)$, denoted by A–Alg–PreSt_B, is defined to be the category of animated A–algebra objects in the category PreSt_B.

Remark 2.14 Another way to define the category A–Alg–PreSt_B is as $\mathrm{Fun}(\mathrm{Alg}_B, \mathrm{ARings}_A)$. However, this is equivalent to the definition considered above since we have natural equivalence of categories

$$\mathrm{Fun}(\mathrm{Alg}_B, \mathrm{ARings}_A) \simeq \mathrm{Fun}(\mathrm{Alg}_B, \mathrm{Fun}_\times(\mathrm{Poly}_A^{\mathrm{op}}, \mathcal{S})) \simeq \mathrm{Fun}_\times(\mathrm{Poly}_A^{\mathrm{op}}, \mathrm{Fun}(\mathrm{Alg}_B, \mathcal{S}))$$
$$\simeq \mathrm{Fun}_\times(\mathrm{Poly}_A^{\mathrm{op}}, \mathrm{PreSt}_B).$$

The middle equivalence uses the fact that product in functor category is calculated termwise; the precise ∞–categorical (dual) assertion can be found in [23, Corollary 5.1.2.3].

Construction 2.15 (cone of a quasi-ideal) In view of Remark 2.14 and Construction 2.3, it follows that, given a quasi-ideal $d : M \to R$ in schemes, the quotient prestack $[R/M]$ (under the additive action of M on the ring scheme R by translation via d) has the structure of a ring prestack. In the context of this paper, we will consider associated ring stacks of such ring prestacks, obtained by fpqc sheafification.

Example 2.16 We will see later that all the examples of quasi-ideals from Section 2.1 have the same cone.

2.3 Affine stacks

We will also use the notion of *affine stacks* due to Toën [29]. Here we will recall its definition and basic properties very briefly, in the language of ∞–categories. To that end, we start by fixing an ordinary base ring B. Let coSCR_B denote the ∞–category of cosimplicial rings over B arising from the simplicial model structure defined in [29, Theorem 2.1.2]; to construct the associated ∞–category from the simplicial model category, one looks at the fibrant simplicial category obtained from the subcategory of fibrant–cofibrant objects inside the given simplicial model category, and applies the simplicial nerve construction, which produces an ∞–category by [23, Proposition 1.1.5.10]. It follows from [23, Corollary 4.2.4.8] that the ∞–category coSCR_B has all small limits and colimits.

Definition 2.17 (affine stacks) An object $\mathcal{Y}$ of PreSt_B is called an *affine stack* over B if there is an object $C \in \mathrm{coSCR}_B$ such that $\mathcal{Y}$ is the restriction of the functor $h_C \colon \mathrm{coSCR}_B \to \mathcal{S}$ corepresented by C along the inclusion $\mathrm{Alg}_B \to \mathrm{coSCR}_B$. The full subcategory of such objects inside PreSt_B is denoted by $\mathrm{AffStacks}_B$.

Remark 2.18 It follows from the definition that the category of affine stacks is stable under small limits; see [29, Proposition 2.2.7]. Also, by [29, Lemma 1.1.2, Proposition 2.2.2], an affine stack is a hypercomplete fpqc sheaf of spaces. The key property of affine stacks that will be useful for us is the fact that taking the derived global section functor induces an equivalence of ∞–categories $\mathrm{AffStacks}_B \simeq \mathrm{coSCR}_B^{\mathrm{op}}$; see [29, Corollary 2.2.3].

Remark 2.19 Even though the definition of the subcategory of affine stacks $\mathrm{AffStacks}_B$ inside PreSt_B a priori depends on the category coSCR_B, the notion of being an affine stack is *intrinsic*: being an affine stack is a property that can be formulated *only* by using the fpqc topology and the category of ordinary rings. See [29, Theorem 2.2.9] for a more precise formulation of this statement using Bousfield localization. A posteriori, the same intrinsic property carries over to the ∞–category coSCR_B, which makes it rather special compared to certain other related categories, such as the ∞–category of derived rings or E_∞–rings.

Example 2.20 An affine scheme is clearly an affine stack. More precisely, the category Aff_B of affine schemes over B embeds fully faithfully inside the category $\mathrm{AffStacks}_B$ of affine stacks over B.

Example 2.21 The stacks $K(\mathbb{G}_a, m)$ for $m \geq 0$ are examples of affine stacks [29, Lemma 2.2.5]. On the other hand, $K(\mathbb{G}_m, m)$ is *not* an affine stack for any $m > 0$. By [29, Corollary 2.4.10], for pointed and connected stacks over a field, being an affine stack is equivalent to the sheaf of all the higher homotopy groups being representable by unipotent affine group schemes (possibly of infinite type).

Remark 2.22 We denote by $\mathrm{St}_B^\wedge$ the ∞–category of hypercomplete fpqc sheaves of spaces (see [23, Section 6.5] for a discussion of hypercomplete ∞–topos). Translating the results [29, Lemma 1.1.2, Proposition 2.2.2, Corollary 2.2.3] into the language of ∞–categories, we obtain a colimit-preserving functor $\mathrm{St}_B^\wedge \to \mathrm{coSCR}_B^{\mathrm{op}}$. There is also a natural colimit-preserving functor $\mathrm{PreSt}_B \to \mathrm{St}_B^\wedge$, and the composite functor denoted by $R\Gamma(\,\cdot\,, \mathcal{O}) \colon \mathrm{PreSt}_B \to \mathrm{coSCR}_B^{\mathrm{op}}$ gives us the "derived global section functor". By construction, $R\Gamma(\,\cdot\,, \mathcal{O}) \colon \mathrm{PreSt}_B \to \mathrm{coSCR}_B^{\mathrm{op}}$ preserves all small colimits. By Definition 2.12 and [23, Lemma 5.1.5.5, Proposition 5.1.5.6], it follows that $R\Gamma(\,\cdot\,, \mathcal{O})$ can be simply described as the left Kan extension of the functor $\mathrm{Aff}_B \to \mathrm{coSCR}_B^{\mathrm{op}}$ (along the inclusion of categories $\mathrm{Aff}_B \to \mathrm{PreSt}_B$) which sends an affine scheme to its underlying ring of global sections. This checks the compatibility of two a priori different ways of defining the derived global section functor.

Remark 2.23 Suppose that $\mathcal{Y}$ is an affine stack over B which is corepresented by $C \in \mathrm{coSCR}_B$ (see Definition 2.17). As noted in Remark 2.18, $\mathcal{Y}$ is a hypercomplete fpqc sheaf of spaces. According to [29, Corollary 2.2.3] and Remark 2.22, we have a natural isomorphism $R\Gamma(\mathcal{Y}, \mathcal{O}) \simeq C$ in coSCR_B. Unwrapping all the definitions and using the equivalence $\mathrm{AffStacks}_B \simeq \mathrm{coSCR}_B^{\mathrm{op}}$, we obtain the categorical implication that the identity functor $\mathrm{coSCR}_B \to \mathrm{coSCR}_B$ is naturally equivalent to the right Kan extension

of the inclusion $\mathrm{Alg}_B \to \mathrm{coSCR}_B$ along itself. Roughly speaking this means that, for any $C \in \mathrm{coSCR}_B$, we have a natural isomorphism

$$C \simeq \Big(\varprojlim_{\substack{C \to A \\ A \text{ is discrete}}} A \Big) \in \mathrm{coSCR}_B.$$

Remark 2.24 The observation in Remark 2.23 regarding right Kan extension implies that if $\mathcal{D}$ is any ∞–category and $F \colon \mathrm{coSCR}_B \to \mathcal{D}$ is a functor that is a right adjoint, then F is naturally equivalent to the right Kan extension of the composite functor $\mathrm{Alg}_B \to \mathrm{coSCR}_B \to \mathcal{D}$ along the inclusion $\mathrm{Alg}_B \to \mathrm{coSCR}_B$.

2.4 Unwinding ring stacks

In this section, we describe how to *unwind*[4] the data of a ring stack to obtain a cohomology theory. This construction is an ∞–categorical enhancement of [24, Example 3.0.1] and we will call this the unwinding of a given ring stack. The construction only uses basic categorical principles such as Kan extensions, and the necessary foundations can be found in [23].

Construction 2.25 (unwinding) We will construct a functor

$$\mathrm{Un} \colon A\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B \to \mathrm{Fun}\big(\mathrm{ARings}_A, \mathrm{CAlg}(D(B))\big).$$

Here $\mathrm{CAlg}(D(B))$ denotes the commutative algebra objects in the derived ∞–category $D(B)$. We think of the objects in the right-hand side as "algebraic cohomology theories".

We begin by noting that by definition $A\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B \simeq \mathrm{Fun}_\times(\mathrm{Poly}_A^{\mathrm{op}}, \mathrm{PreSt}_B)$. By Kan extension, there is a derived global section functor $R\Gamma \colon \mathrm{PreSt}_B \to \mathrm{CAlg}(D(B))^{\mathrm{op}}$. By composition, we get a functor

$$A\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B^{\mathrm{op}} \to \mathrm{Fun}\big(\mathrm{Poly}_A, \mathrm{CAlg}(D(B))\big).$$

Now we can perform a left Kan extension along the inclusion $\mathrm{Poly}_A \to \mathrm{ARings}_A$ to obtain the desired unwinding functor

$$\mathrm{Un} \colon A\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B^{\mathrm{op}} \to \mathrm{Fun}\big(\mathrm{ARings}_A, \mathrm{CAlg}(D(B))\big).$$

Example 2.26 When $A = B$ and $\mathcal{Y} \in \mathrm{PreSt}_B$ is taken to be the ring scheme $\mathbb{G}_{a,B}$, the functor $\mathrm{Un}(\mathbb{G}_{a,B})$ is simply the forgetful functor $\mathrm{ARings}_A \to \mathrm{CAlg}(D(B))$.

Below we will study compatibility of the unwinding construction with restriction of scalars. More precisely, let $\mathcal{Y} \in A\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B$. Let $A' \to A$ be a map of discrete rings. Then there is an obvious functor

$$\mathrm{res} \colon A\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B \to A'\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B.$$

Let $\mathcal{Y}' := \mathrm{res}(\mathcal{Y}) \in A'\text{–}\mathrm{Alg}\text{–}\mathrm{PreSt}_B$. Applying the unwinding construction, we obtain two functors $\mathrm{Un}(\mathcal{Y}) \colon \mathrm{ARings}_A \to \mathrm{CAlg}(D(B))$ and $\mathrm{Un}(\mathcal{Y}') \colon \mathrm{ARings}_{A'} \to \mathrm{CAlg}(D(B))$. Note that we also have a natural functor (given by the derived tensor product) $L \colon \mathrm{ARings}_{A'} \to \mathrm{ARings}_A$. In this setup, we have the following compatibility:

---

[4]A similar construction has been used by Bhatt in [4], under the name transmutation.

Proposition 2.27 *We have* $\mathrm{Un}(\mathcal{Y}) \circ L \simeq \mathrm{Un}(\mathcal{Y}')$ *in* $\mathrm{Fun}\big(\mathrm{ARings}_A, \mathrm{CAlg}(D(B))\big)$.

Proof Since L is obtained by left Kan extension of the composite functor $\mathrm{Poly}_{A'} \xrightarrow{\ell} \mathrm{Poly}_A \to \mathrm{ARings}_A$, it would be enough to prove $\mathrm{Un}(\mathcal{Y}) \circ \ell \simeq \mathrm{Un}(\mathcal{Y}')$ in $\mathrm{Fun}\big(\mathrm{Poly}_A, \mathrm{CAlg}(D(B))\big)$. By Construction 2.25, $\mathcal{Y}$ is classified by a functor $U \colon \mathrm{Poly}_A^{\mathrm{op}} \to \mathrm{PreSt}_B$ and $\mathcal{Y}'$ is classified by $U' \colon \mathrm{Poly}_{A'}^{\mathrm{op}} \to \mathrm{PreSt}_B$; for our purpose, it would be enough to prove that $U \circ \ell^{\mathrm{op}} \simeq U'$. By Remark 2.14, it would be enough to prove that the restriction of scalar functor $\mathrm{ARings}_A \to \mathrm{ARings}_{A'}$ is induced by ℓ^{op} under the identifications $\mathrm{ARings}_A \simeq \mathrm{Fun}_\times(\mathrm{Poly}_A^{\mathrm{op}}, \mathcal{S})$ and $\mathrm{ARings}_{A'} \simeq \mathrm{Fun}_\times(\mathrm{Poly}_{A'}^{\mathrm{op}}, \mathcal{S})$. But that follows from adjunction. $\qquad\square$

Notation 2.28 If k is a perfect field of characteristic p and $\mathcal{Y}$ is a $W_n(k)$–algebra stack for $1 \le n \le \infty$, then we will use $\mathcal{Y}^{(1)}$ to denote the $W_n(k)$–algebra stack obtained by restriction of scalars along the Witt vector Frobenius $W_n(k) \to W_n(k)$; see Proposition 2.27.

Remark 2.29 Here the Frobenius twist $\mathcal{Y}^{(1)}$ of a stack $\mathcal{Y}$ will not play an important role, because we always work over a perfect field and are interested in the question of endomorphisms of the stacks. Since it also does not change the underlying stack, for the most part we will ignore this Frobenius twist.

Example 2.30 Proposition 2.27 shows that the Frobenius twisted forgetful functor

$$R \mapsto R^{(1)} := R \otimes_{k,\mathrm{Frob}} k$$

from $\mathrm{ARings}_k \to \mathrm{CAlg}(D(k))$ is the unwinding of $\mathbb{G}_{a,k}^{(1)}$. The relative Frobenius $R^{(1)} \to R$ can be obtained by unwinding the map of k–algebra stacks $\mathbb{G}_{a,k} \to \mathbb{G}_{a,k}^{(1)}$ induced by the Frobenius.

2.5 De Rham cohomology via unwinding

In this section, we will describe how to use the unwinding construction to recover de Rham or crystalline cohomology functors. To this end, let $n, m \ge 1$ be two arbitrary positive integers and let p be a fixed prime. Further, we fix a perfect ring k of characteristic p. Let $W_r(k)$ denote the ring of r–truncated Witt vectors. Using crystalline cohomology, or more precisely its derived variant (see Definition 2.31 below), one obtains certain functors denoted by

$$\mathrm{dR}_{m,n} \colon \mathrm{ARings}_{W_n(k)} \to \mathrm{CAlg}\big(D(W_m(k))\big),$$

which we loosely still call de Rham cohomology functors and specify the n and m. To define them, one really needs to use a deformation of the de Rham cohomology functor, ie the crystalline cohomology functors.

The following essentially already appeared in [7, Section 10.2; 9, Section 8.2].

Definition 2.31 Let P be a finitely generated polynomial $W_n(k)$–algebra. Then define $\mathrm{dR}_{m,n}(P) :=$ $\mathrm{R}\Gamma_{\mathrm{crys}}(P_0/W_m(k))$, where P_0 denotes the mod p reduction of P. We denote by $\mathrm{dR}_{m,n}$ the left Kan extension of the above functor from finitely generated polynomials to all animated $W_n(k)$–algebras which takes values in $\mathrm{CAlg}\big(D(W_m(k))\big)$.

We use this notation as we believe the crystalline cohomology is secretly a disguise of derived de Rham cohomology; see [3, Proposition 3.27; 21, Proposition 2.11] for instances of this perspective. Our goal is to describe $\mathrm{dR}_{m,n}$ as the unwinding of a certain object in $W_n(k)\text{–Alg–PreSt}_{W_m(k)}$.

Definition 2.32 Let W denote the ring scheme over $\mathrm{Spec}(W_m(k))$ underlying the p–typical Witt vectors. Using the Artin–Hasse homomorphism $W(k) \to W(W(k))$, one can view W as a $W(k)$–algebra scheme. Then $d : W^{(1)} \xrightarrow{\times p} W^{(1)}$ defines a quasi-ideal in schemes. By considering its cone, one obtains a k–algebra stack over $\mathrm{Spec}(W_m(k))$, which can be regarded as a $W_n(k)$–algebra stack over $\mathrm{Spec}(W_m(k))$ via the natural map $W_n(k) \twoheadrightarrow k$. We denote the resulting $W_n(k)$–algebra stack over $\mathrm{Spec}(W_m(k))$ by $\mathbb{A}^{1,\mathrm{dR}}_{m,n}$. When n is fixed, we will use $\mathbb{A}^{1,\mathrm{dR}}_{B}$ to denote the pullback of $\mathbb{A}^{1,\mathrm{dR}}_{m,n}$ to $\mathrm{Spec}\,B$ for a $W_m(k)$–algebra B.

Remark 2.33 The above definition gives a generalization of the definition of $\mathbb{A}^{1,\mathrm{dR}}$ as an $\mathbb{F}_p$–algebra stack due to Drinfeld to the more general case of an arbitrary perfect ring k. To do this, one crucially needs to use the Artin–Hasse natural transformation $W(\cdot) \to W(W(\cdot))$. One can abstractly construct this natural transformation by realizing the functor W as a right adjoint to the inclusion of the category of delta rings inside all rings.

Proposition 2.34 *The stack underlying $\mathbb{A}^{1,\mathrm{dR}}_{m,n}$ is an affine stack.*

Proof Indeed, the stack underlying $\mathbb{A}^{1,\mathrm{dR}}_{m,n}$ is obtained by taking the cone of $d : W \xrightarrow{\times p} W$, which is the same as the fiber of the induced map $BW \to BW$. Since affine stacks are closed under limits, it would be enough to show that BW is an affine stack. This follows from the proof of [26, Proposition 3.2.7]. Let us give a rough sketch of their argument. Let W_n denote the ring scheme underlying n–truncated p–typical Witt vectors. Using that certain obstructions vanish, one first argues that $BW \simeq \varprojlim BW_n$. Therefore, it is enough to prove that BW_n is an affine stack for all n. To do so, one argues by induction on n. Using the short exact sequence

$$0 \to \mathbb{G}_a \to W_{n+1} \to W_n \to 0,$$

one sees that BW_{n+1} is classified by a map $BW_n \to K(\mathbb{G}_a, 2)$. More precisely, we have a fiber sequence

$$
\begin{array}{ccc}
BW_{n+1} & \longrightarrow & * \\
\downarrow & & \downarrow \\
BW_n & \longrightarrow & K(\mathbb{G}_a, 2)
\end{array}
$$

Since the stacks $K(\mathbb{G}_a, m)$ are affine stacks for $m \geq 0$, we are done by induction. $\qquad\square$

Remark 2.35 The above argument can be modified to more generally show that $K(W, m)$ is an affine stack for all $m \geq 0$. Consequently, one can show that the abelian group stack $\mathbb{A}^{1,\mathrm{dR}}[m]$ is also an affine stack for all $m \geq 0$. We have $R\Gamma_{\mathrm{dR}}(K(\mathbb{G}_a, m)) \simeq R\Gamma(\mathbb{A}^{1,\mathrm{dR}}[m], \mathcal{O})$ for all $m \geq 0$.

Proposition 2.36 [6, Remark 7.9; 25] *We have a natural isomorphism $\mathrm{Un}(\mathbb{A}^{1,\mathrm{dR}}_{m,n}) \simeq \mathrm{dR}_{m,n}$.*

Proof By Proposition 2.27, the proof reduces to $n = 1$. Further, by [24, Theorem 1.1.1], one can reduce to $m = 1$. Let us now explain the proof of the natural isomorphism $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}}) \simeq \mathrm{dR}_{1,1}$. By construction of the unwinding functor, it would be enough to show that the restricted functors $\mathrm{dR}_{1,1} \colon \mathrm{Poly}_k \to \mathrm{CAlg}(D(k))$ and $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}}) \colon \mathrm{Poly}_k \to \mathrm{CAlg}(D(k))$ are naturally isomorphic. Note that we have a natural functor $\mathrm{coSCR}_k \to \mathrm{CAlg}(D(k))$ of ∞–categories, and by construction $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})$ lifts to give a functor, still denoted by $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}}) \colon \mathrm{Poly}_k \to \mathrm{coSCR}_k$. By quasisyntomic descent, $\mathrm{dR}_{1,1}$ also lifts to give a functor, still denoted by $\mathrm{dR}_{1,1} \colon \mathrm{Poly}_k \to \mathrm{coSCR}_k$. It would be enough to prove that these two functors are naturally isomorphic.

By considering gr^0 of the Hodge filtration on de Rham cohomology and quasisyntomic descent, there is a natural arrow $\mathrm{dR}_{1,1} \to \iota$ in the category $\mathrm{Fun}(\mathrm{Poly}_k, \mathrm{coSCR}_k)$, where $\iota \colon \mathrm{Poly}_k \to \mathrm{coSCR}_k$ denotes the natural inclusion functor.

Note that the derived global sections of $\mathbb{A}_{1,1}^{1;\mathrm{dR}}$ agree with $\mathrm{dR}_{1,1}(k[x])$ in coSCR_k. For this, one can use the identification $\mathrm{Cone}(\mathbb{G}_a^\sharp \to \mathbb{G}_a) \simeq \mathrm{Cone}(W \xrightarrow{\times p} W)$ and the Čech–Alexander complex. Since, by Proposition 2.34, $\mathbb{A}_{1,1}^{1;\mathrm{dR}}$ is an affine stack, it follows that the functors $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}}) \colon \mathrm{Poly}_k \to \mathrm{coSCR}_k$ and $\mathrm{dR}_{1,1} \colon \mathrm{Poly}_k \to \mathrm{coSCR}_k$ preserve finite coproducts. In order to check that they are naturally isomorphic, it is enough to do so for the functors $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})' \colon \mathrm{ARings}_k \to \mathrm{coSCR}_k$ and $\mathrm{dR}_{1,1}' \colon \mathrm{ARings}_k \to \mathrm{coSCR}_k$ obtained by left Kan extension along $\mathrm{Poly}_k \to \mathrm{ARings}_k$.

By [23, Proposition 5.5.8.15], the functors $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})'$ and $\mathrm{dR}_{1,1}'$ both preserve small colimits. Similarly, by left Kan extension, $\iota \colon \mathrm{Poly}_k \to \mathrm{coSCR}_k$ extends to a colimit-preserving functor $\iota' \colon \mathrm{ARings}_k \to \mathrm{coSCR}_k$. By the adjoint functor theorem, all of these functors have right adjoints. Let $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})'^{\mathcal{R}}$, $\mathrm{dR}_{1,1}'^{\mathcal{R}}$ and $\iota'^{\mathcal{R}}$ denote the right adjoints to $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})'$, $\mathrm{dR}_{1,1}'$ and ι', respectively. It would be enough to prove that $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})'^{\mathcal{R}} \simeq \mathrm{dR}_{1,1}'^{\mathcal{R}}$.

Let $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})_\circ'^{\mathcal{R}}$, $\mathrm{dR}_{1,1\circ}'^{\mathcal{R}}$ and $\iota_\circ'^{\mathcal{R}}$ denote the restrictions of the functors $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})'^{\mathcal{R}}$, $\mathrm{dR}_{1,1}'^{\mathcal{R}}$ and $\iota'^{\mathcal{R}}$, respectively, along the inclusion of categories $\mathrm{Alg}_k \to \mathrm{coSCR}_k$. For our purpose, by considering right Kan extensions as explained in Remark 2.24, it would be enough to prove that $\mathrm{Un}(\mathbb{A}_{1,1}^{1;\mathrm{dR}})_\circ'^{\mathcal{R}} \simeq \mathrm{dR}_{1,1\circ}'^{\mathcal{R}}$. Note that they are both functors from Alg_k to ARings_k. Further, for an $S \in \mathrm{Alg}_k$, we have $\iota_\circ'^{\mathcal{R}}(S) = S$, which identifies with the S–valued points of the ring scheme $\mathbb{G}_a$. Thus we have a natural arrow $\iota_\circ'^{\mathcal{R}} \simeq \mathbb{G}_a \to \mathrm{dR}_{1,1\circ}'^{\mathcal{R}}$ in $\mathrm{Fun}(\mathrm{Alg}_k, \mathrm{ARings}_k)$, where $\mathbb{G}_a$ is viewed as an object of $\mathrm{Fun}(\mathrm{Alg}_k, \mathrm{ARings}_k)$ by considering its functor of points. We note the following lemma:

Lemma 2.37 *The fiber F of the map $\mathbb{G}_a \to \mathrm{dR}_{1,1\circ}'^{\mathcal{R}}$ identifies with the $\mathbb{G}_a$–module scheme $\mathbb{G}_a^\sharp$.*

Proof To see this, we note that $\mathrm{dR}_{1,1\circ}'^{\mathcal{R}}$ can be viewed as a ring stack whose underlying stack, by construction, is the affine stack corresponding to the object $\mathrm{dR}_{1,1}'(k[x]) \in \mathrm{coSCR}_k$. Therefore, the stack underlying F is given by the affine stack corresponding to the cosimplicial ring obtained by the pushout $k \sqcup_{\mathrm{dR}_{1,1}'(k[x])} k[x]$ in coSCR_k. Since $\mathbb{A}_{1,1}^{1;\mathrm{dR}}$ is an affine stack and $R\Gamma(\mathbb{A}_{1,1}^{1;\mathrm{dR}}, \mathcal{O}) \simeq \mathrm{dR}_{1,1}'(k[x])$, it follows that $k \sqcup_{\mathrm{dR}_{1,1}'(k[x])} k[x] \simeq R\Gamma(\mathbb{G}_a^\sharp, \mathcal{O}) \simeq D_x(k[x])$, where $D_x(k[x])$ denotes the divided power envelope of $k[x]$ at the ideal (x). In particular, the pushout is a discrete ring, and the stack underlying F is an

affine scheme. Let $dR(k[x])$ denote the object of $CAlg(D(k))$ underlying $dR'_{1,1}(k[x])$. Then there is a natural map

$$k \otimes_{dR(k[x])} k[x] \to k \sqcup_{dR'_{1,1}(k[x])} k[x]$$

in $CAlg(D(k))$. We have an isomorphism $k \otimes_{dR(k[x])} k[x] \simeq dR_{k/k[x]}$, where $dR_{k/k[x]}$ denotes derived de Rham cohomology. By [3, Lemma 3.29], it follows that $dR_{k/k[x]} \simeq D_x(k[x])$ and the natural map above is an isomorphism. We note that $Spec(k \otimes_{dR(k[x])} k[x]) \simeq Spec(dR_{k/k[x]})$ also has the structure of a group scheme, where the multiplication is induced by functoriality of $dR_{k/(\cdot)}$ along the map $k[x] \to k[x] \otimes_k k[x]$ given by $x \mapsto x \otimes 1 + 1 \otimes x$. Moreover, $Spec(dR_{k/k[x]})$ has the structure of a $\mathbb{G}_a$–equivariant group scheme, where the $\mathbb{G}_a$–action is given by the map $k[x] \simeq dR_{k[x]/k[x]} \to dR_{k/k[x]} \otimes_k k[x] \simeq D_x(k[x]) \otimes_k k[x]$ which is induced by functoriality of derived de Rham cohomology applied to the diagram

$$
\begin{array}{ccc}
k[x] & \xrightarrow{\ \ x \to x\ \ } & k[x] \\
{\scriptstyle x \to x \otimes x} \downarrow & & \downarrow {\scriptstyle x \to 0} \\
k[x] \otimes k[x] & \xrightarrow{x \otimes 1 \to 0, 1 \otimes x \to x} & k[x]
\end{array}
$$

Using the explicit description of the induced maps, one explicitly verifies that $Spec(dR_{k/k[x]})$ is naturally isomorphic to $\mathbb{G}_a^\sharp$ as a $\mathbb{G}_a$–module scheme. Further, by applying functoriality along the diagrams mentioned earlier, we see that the map of schemes $F \to Spec(dR_{k/k[x]})$ induced by the natural map $k \otimes_{dR(k[x])} k[x] \to k \sqcup_{dR'_{1,1}(k[x])} k[x]$ above is actually a $\mathbb{G}_a$–equivariant map of group schemes. Since we have already noted that $k \otimes_{dR(k[x])} k[x] \to k \sqcup_{dR'_{1,1}(k[x])} k[x]$ is an isomorphism, this shows that F is indeed isomorphic to $\mathbb{G}_a^\sharp$ as a $\mathbb{G}_a$–module scheme, as desired. □

Now we have obtained a natural map $\mathbb{A}^{1,dR}_{1,1} \simeq Cone(\mathbb{G}_a^\sharp \to \mathbb{G}_a) \to dR'_{1,1}{}^\mathcal{R}_\circ$ of k–algebra stacks, ie as objects in the category $Fun(Alg_k, ARings_k)$. We have already noted that their underlying stacks are isomorphic. Thus we obtain an isomorphism $\mathbb{A}^{1,dR}_{1,1} \simeq dR'_{1,1}{}^\mathcal{R}_\circ$. Since the stack underlying $\mathbb{A}^{1,dR}_{1,1}$ is an affine stack, it follows that $Un(\mathbb{A}^{1,dR}_{1,1})'^\mathcal{R}_\circ \simeq \mathbb{A}^{1,dR}_{1,1}$ as objects of $Fun(Alg_k, ARings_k)$. This constructs the isomorphism $Un(\mathbb{A}^{1,dR}_{1,1})'^\mathcal{R}_\circ \simeq dR'_{1,1}{}^\mathcal{R}_\circ$, which finishes the proof. □

The following fact was used in the above proof, which uses compatibility of two models of the k–algebra stack $\mathbb{A}^{1,dR}$ over $Spec(W(k))$.

Proposition 2.38 [14, 3.5.1] *There is an isomorphism of k–algebra stacks over $Spec(W(k))$:*

$$Cone(\mathbb{G}_a^\sharp \to \mathbb{G}_a) \simeq Cone(W^{(1)} \xrightarrow{\times p} W^{(1)}).$$

The k–algebra structure on the source comes from the natural maps $W(k) \to \mathbb{G}_a$ and $W(k) \xrightarrow{1 \mapsto p - V(1)} W[F]$. To see that the two underlying abelian group stacks are the same, notice that we always have $FV = p$ on the p–typical Witt ring, and hence we get a factorization

$$
\begin{array}{ccc}
W^{(1)} & \xrightarrow{\ \ \times p\ \ } & W^{(1)} \\
& {\scriptstyle V} \searrow \quad \nearrow {\scriptstyle F} & \\
& W &
\end{array}
$$

One then applies the octahedral axiom to the above triangle. The fact that it induces an algebra isomorphism can be seen, using the fact that F is an algebra homomorphism. Said differently, one pulls back the quasi-ideal $W^{(1)} \xrightarrow{\times p} W^{(1)}$ along $W \xrightarrow{F} W^{(1)}$ to build the intermediate model relating the above two models.

Remark 2.39 There is a natural map of k–algebra stacks $\mathbb{G}_a \to \mathbb{A}^{1,\mathrm{dR}}$ whose unwinding provides a natural transformation $\mathrm{dR}(S) \to S$, which corresponds to the natural projection onto the gr^0 of the Hodge filtration on de Rham cohomology. There is also a natural map $\mathbb{A}^{1,\mathrm{dR}} \to \pi_0(\mathbb{A}^{1,\mathrm{dR}}) = \mathbb{G}_a^{(1)}$ of k–algebra stacks which unwinds to the natural transformation $S^{(1)} \to \mathrm{dR}(S)$ induced by Fil^0 of the conjugate filtration; see Proposition 4.20.

Now we will see that the quasi-ideal $\mathbb{G}_a^{\mathrm{perf},\sharp} \to \mathbb{G}_a^{\mathrm{perf}}$ that appears in [24, Proposition 4.0.11] gives a third model of the k–algebra stack $\mathbb{A}^{1,\mathrm{dR}}$ over $\mathrm{Spec}(W(k))$; see also [13]. First, we will make some preparations. Below we always fix a positive integer m.

Lemma 2.40 On the fpqc site of $\mathbb{Z}/p^m$, we have $R \varprojlim_F W \simeq \varprojlim_F W$, which is representable by an affine scheme. Moreover, its functor of points can be described as $B \mapsto W(B^\flat)$.

We denote the affine scheme representing $\varprojlim_F W$ by W^{perf}. This scheme can be given an $W(k)$–algebra scheme structure when viewed over $\mathrm{Spec}(W_m(k))$.

Proof The first assertion follows from [10, Example 3.1.7 and Proposition 3.1.10] and the fact that F on W is faithfully flat. The inverse limit of affine schemes is again affine. For the last claim, we consider the following diagram of fpqc sheaves as a pro-object:

$$
\begin{array}{ccccc}
\cdots \longrightarrow & W_3 & \xrightarrow{F} & W_2 & \xrightarrow{F} & W_1 \\
& \uparrow{\scriptstyle R} & & \uparrow{\scriptstyle R} & & \uparrow{\scriptstyle R} \\
\cdots \longrightarrow & W_4 & \xrightarrow{F} & W_3 & \xrightarrow{F} & W_2 \\
& \uparrow{\scriptstyle R} & & \uparrow{\scriptstyle R} & & \uparrow{\scriptstyle R} \\
\cdots \longrightarrow & W_5 & \xrightarrow{F} & W_4 & \xrightarrow{F} & W_3 \\
& \uparrow{\scriptstyle R} & & \uparrow{\scriptstyle R} & & \uparrow{\scriptstyle R} \\
& \vdots & & \vdots & & \vdots
\end{array}
$$

Taking the limit vertically and then horizontally gives us $\varprojlim_F W$. Next we take limit horizontally and then vertically instead. Taking limits horizontally, we obtain the sheaf that sends B to $\varprojlim_F W_r(B)$, which is canonically identified with $W(B^\flat)$ by [8, Lemma 3.2] (with π in loc. cit. being p). The vertical map R is actually an isomorphism now, also by [8, Lemma 3.2]. This gives $\varprojlim_F W(B) \simeq W(B^\flat)$, as desired. $\square$

Recall that F on W induces a map $\mathbb{A}^{1,\mathrm{dR}} \to \mathrm{Frob}_{k,*} \mathbb{A}^{1,\mathrm{dR}}$ of k–algebra stacks which we will again denote by F. We may untwist the Frobenius using the inverse of the Frobenius on k on the source of this map. Therefore we get a k–algebra structure on the stack $\varprojlim_F(\mathbb{A}^{1,\mathrm{dR}})$.

Lemma 2.41 *We have an isomorphism of k–algebra stacks $\mathbb{G}_a^{\mathrm{perf}} \simeq \varprojlim_F (\mathbb{A}^{1,\mathrm{dR}})$ over $\mathrm{Spec}(W_m(k))$.*

Proof By Lemma 2.40, we see that

$$R \varprojlim_F (\mathbb{A}^{1,\mathrm{dR}}) = \mathrm{Cone}\big(R \varprojlim_F W \xrightarrow{\times p} R \varprojlim_F W \big) = \mathrm{Cone}(W^{\mathrm{perf}} \xrightarrow{\times p} W^{\mathrm{perf}}),$$

and its functor of points is given by $B \mapsto B^{\flat}$. Hence $\varprojlim_F (\mathbb{A}^{1,\mathrm{dR}})$ is isomorphic to $\mathbb{G}_a^{\mathrm{perf}}$ as a k–algebra stack (and in fact is a scheme). $\qquad\square$

Therefore we get a map of k–algebra stacks $\mathbb{G}_a^{\mathrm{perf}} \to \mathbb{A}^{1,\mathrm{dR}}$ over $\mathrm{Spec}(W_m(k))$.

Lemma 2.42 *The map of (k–algebra) stacks $f : \mathbb{G}_a^{\mathrm{perf}} \to \mathbb{A}^{1,\mathrm{dR}}$ is faithfully flat.*

Proof We look at the diagram of k–algebra stacks

$$
\begin{array}{ccc}
W^{\mathrm{perf}} & \longrightarrow & \mathbb{G}_a^{\mathrm{perf}} \\
& \searrow \quad \swarrow & \\
& \mathbb{A}^{1,\mathrm{dR}} &
\end{array}
$$

and observe that the horizontal and the left arrow are faithfully flat, and hence the right arrow is faithfully flat as well. $\qquad\square$

Let K be the quasi-ideal in $\mathbb{G}_a^{\mathrm{perf}}$ given by the kernel of f. Then Lemma 2.42 implies that f gives rise to an isomorphism of k–algebra stacks $\mathrm{Cone}(K \to \mathbb{G}_a^{\mathrm{perf}}) \simeq \mathbb{A}^{1,\mathrm{dR}}$. This is what we called the third model of $\mathbb{A}^{1,\mathrm{dR}}$; to complete the description, it remains to understand the quasi-ideal K.

Proposition 2.43 *K is isomorphic to $\mathbb{G}_a^{\mathrm{perf},\sharp}$ as a quasi-ideal in $\mathbb{G}_a^{\mathrm{perf}}$. In particular, as k–algebra stacks, $\mathrm{Cone}(\mathbb{G}_a^{\mathrm{perf},\sharp} \to \mathbb{G}_a^{\mathrm{perf}}) \simeq \mathbb{A}^{1,\mathrm{dR}}$.*

Proof This assertion follows from applying the (derived) crystalline cohomology functor $R\Gamma_{\mathrm{crys}}$ to the pushout diagram

$$
\begin{array}{ccc}
k & \longrightarrow & k[x^{1/p^{\infty}}]/x \\
{\scriptstyle x \mapsto 0} \uparrow & & \uparrow \\
k[x] & \longrightarrow & k[x^{1/p^{\infty}}]
\end{array}
$$

and noting that global sections of $\mathbb{G}_a^{\mathrm{perf},\sharp}$ recover $R\Gamma_{\mathrm{crys}}(k[x^{1/p^{\infty}}]/x)$ and $R\Gamma_{\mathrm{crys}}$ preserves the pushout diagram. $\qquad\square$

Remark 2.44 Using the above methods, let us sketch a quick proof of a result due to Bhatt, Lurie and Mathew [7, Proposition 10.3.1]; see also [24, Proposition 4.0.7]. Under Proposition 2.36, Example 2.26 and Remark 2.39, the assertion amounts to studying endomorphisms of $\mathbb{A}^{1,\mathrm{dR}}$ respecting the natural map $\mathbb{G}_a \to \mathbb{A}^{1,\mathrm{dR}}$. By Proposition 2.38, it is enough to show that the quasi-ideal $\mathbb{G}_a^{\sharp} \to \mathbb{G}_a$ has no nontrivial endomorphism as a quasi-ideal in $\mathbb{G}_a$. This follows directly from graded Cartier duality [24, Section 2.4].

Remark 2.45 The definition of $\mathbb{A}^{1,\mathrm{dR}}$ as a k–algebra stack differs from $\mathrm{Cone}(W \xrightarrow{\times p} W)$ by a Frobenius twist. Indeed, the latter unwinds to Hodge–Tate cohomology (or a suitable base change of prismatic cohomology) [11] which is the Frobenius descent of de Rham cohomology (or crystalline cohomology); see Proposition 2.27.

3   Endomorphisms of de Rham cohomology, I

The quasisyntomic descent technique introduced in [9] is a powerful tool in calculating endomorphisms of de Rham cohomology functors in various settings. We will illustrate them in this section.

Let $A \to B$ be a map of derived p–complete rings with bounded p^∞–torsion. In this section, we consider the functor that, for a derived p–complete A–algebra R, is defined by

$$(\mathrm{dR}\widehat{\otimes}_A B)(R) := \mathrm{dR}_{R/A}\widehat{\otimes}_A B \in \mathrm{CAlg}(D(B)),$$

where $\mathrm{dR}_{R/A}$ denotes the p–adic derived de Rham complex of R relative to A and $\widehat{\otimes}$ denotes the derived p–completed tensor product. If $B = A$, we simply denote the functor by dR.

We are interested in the space of endomorphisms of this functor, viewed (by left Kan extension) as an object in the ∞–category of functors from the ∞–category of derived p–complete animated rings to $\mathrm{CAlg}(D(B))$. Let qSyn_A denote the small quasisyntomic site of A which consists of algebras that are quasisyntomic over A and the covers given by quasisyntomic covers; see [9, Section 4.2].

Proposition 3.1 (see [9, Example 5.12]) *The functor $\mathrm{dR}\widehat{\otimes}_A B$, when restricted to qSyn_A, defines a quasisyntomic sheaf.*

Proof It suffices to check this after going derived modulo p, so we are reduced to checking the following: given $R \to S$ a faithfully flat quasisyntomic map of algebras in qSyn_A with Čech nerve $S^\bullet$, there is an isomorphism

$$\mathrm{dR}_{R/A} \otimes_A B/p \simeq \lim(\mathrm{dR}_{S^\bullet/A} \otimes_A B/p),$$

where B/p is the animated ring $B \otimes_{\mathbb{Z}} \mathbb{F}_p$. By base change of derived de Rham cohomology, this is equivalent to showing

$$\mathrm{dR}_{(R \otimes_A B/p)/(B/p)} \simeq \lim(\mathrm{dR}_{(S^\bullet \otimes_A B/p)/(B/p)}).$$

See [19, pages 33–35] for a discussion of derived de Rham cohomology of maps of animated rings. To prove the above isomorphism, we employ the conjugate filtration [19, Construction 2.3.12] (with base ring $\mathbb{F}_p$). The conjugate filtration is exhaustive and uniformly bounded above by -1, and hence it suffices to prove that its graded pieces satisfy similar quasisyntomic descent. Using the description of graded pieces of conjugate filtration, we are finally reduced to showing

$$\bigwedge_A^i \mathbb{L}_{R/A} \otimes_A \varphi_*(B/p) \simeq \lim\big(\bigwedge_{S^\bullet}^i \mathbb{L}_{S^\bullet/A} \otimes_A \varphi_*(B/p)\big).$$

Here $\varphi_*(B/p)$ expresses the A–module structure on B/p which is given by $A \to B \to B/p \xrightarrow{\varphi} B/p$. Proposition 3.2 finishes the proof. $\square$

Proposition 3.2 (flat descent for "tensored" cotangent complex)   *Fix a base ring A. For each $n \geq 0$ and an object $M \in D(A)$, the functor $R \mapsto \bigwedge_R^n \mathbb{L}_{R/A} \otimes_A M$ is an fpqc sheaf with values in the ∞–category $D(A)$.*

Proof   One simply runs through the proof of [9, Theorem 3.1] and sees that it works in this generality. For convenience of the reader, let us illustrate the proof when $n = 1$. Let $R \to S$ be a faithfully flat map of A–algebras with Čech nerve $S^\bullet$. Using the transitivity triangle associated to $A \to R \to S^\bullet$ and applying the exact functor $(\,\cdot\,) \otimes_A M$, we get a cosimplicial exact triangle

$$\mathbb{L}_{R/A} \otimes_R S^\bullet \otimes_A M \to \mathbb{L}_{S^\bullet/A} \otimes_A M \to \mathbb{L}_{S^\bullet/R} \otimes_A M.$$

We are therefore reduced to showing:

- The map $R \to S^\bullet$ induces an isomorphism $\mathbb{L}_{R/A} \otimes_A M \to \lim \mathbb{L}_{R/A} \otimes_A M \otimes_R S^\bullet$.
- $\lim \mathbb{L}_{S^\bullet/R} \otimes_A M = 0$.

The first item follows from fpqc descent along $R \to S$ by considering $\mathbb{L}_{R/A} \otimes_A M \in D(R)$. The second is proved via a few reduction steps. By the convergence of the Postnikov filtration, it is enough to show that $\lim \pi_i(\mathbb{L}_{S^\bullet/R} \otimes_A M) \simeq 0$ in $D(R)$ for an arbitrary $i \in \mathbb{Z}$, which will be fixed from now. Again, by faithfully flat descent, it suffices to check that $(\lim \pi_i(\mathbb{L}_{S^\bullet/R} \otimes_A M)) \otimes_R S \simeq \lim(\pi_i(\mathbb{L}_{S^\bullet/R} \otimes_A M) \otimes_R S) \simeq \lim \pi_i(\mathbb{L}_{S^\bullet/R} \otimes_R S \otimes_A M) \simeq 0$. Let $S \to T^\bullet$ denote the base change of $R \to S^\bullet$ along $R \to S$. By base change for cotangent complex, we need to show that $\lim \pi_i(\mathbb{L}_{T^\bullet/S} \otimes_A M) \simeq 0$. Since $S \to T^\bullet$ is the Čech nerve of the map $S \to S \otimes_R S$, which admits a section, it follows that $S \to T^\bullet$ is a homotopy equivalence of cosimplicial S–algebras. Now we observe that $F := \pi_i(\mathbb{L}_{(\cdot)/S} \otimes_A M)$ is a functor from the category of S–algebras to the category of abelian groups. Therefore the cosimplicial abelian group $F(T^\bullet)$ is homotopy equivalent to $F(S)$. Since $F(S) \simeq 0$, we obtain $\lim \pi_i(\mathbb{L}_{T^\bullet/S} \otimes_A M) \simeq 0$, as desired. $\qquad \square$

As a consequence, let us record a result that says that the space of endomorphisms is actually discrete, ie the homotopy groups in degrees above zero are trivial for every choice of basepoints.

Lemma 3.3   *The space of endomorphisms $\mathrm{End}(\mathrm{dR}\widehat{\otimes}_A B)$ is discrete.*

Proof   First observe that $\mathrm{dR}\widehat{\otimes}_A B$ is left Kan extended from its restriction to the category of p–completely finitely generated polynomial A–algebras. Hence the restricted functor has the same space of endomorphisms. Since our functor $\mathrm{dR}\widehat{\otimes}_A B$ is a sheaf on the quasisyntomic site of A and since p–completed polynomial A–algebras are quasisyntomic over A, restricting our functor to the full subcategory of A–algebras consisting of algebras that are quasisyntomic over A again computes the same endomorphism space. Recall that, since the quasisyntomic site of A admits a basis consisting of large quasisyntomic A–algebras (see [11, Definition 15.1]), we may restrict our (base-changed) de Rham cohomology functor to this basis and compute the space of endomorphisms there. But now the values of the de Rham cohomology functor are p–completely flat A–algebras, and hence the base-changed de Rham cohomology functor has values which are discrete B–algebras [9, Lemma 4.6]. Consequently the space of endomorphisms is discrete. $\qquad \square$

If $\mathrm{Spf}(A)$ has a disconnection, then the space of endomorphisms will be the product of endomorphism spaces on each subset giving rise to the disconnection. Hence, without loss of generality, let us only treat those A with connected formal spectrum. The following simple lemma will be used later, so let us record it here:

Lemma 3.4 *Let A_0 be an idempotent-free $\mathbb{F}_p$–algebra. Let q be a power of p. If every element $a \in A_0$ satisfies $a^q = a$, then A_0 is a subfield inside $\mathbb{F}_q$.*

In the rest of this section we will compute the space of endomorphisms in two cases:

Case I A is the Witt ring of an idempotent-free characteristic-p perfect algebra k and $B = A$.

Case II A is a perfect $\mathbb{F}_p$–algebra and B is an arbitrary A–algebra.

Building on the method of [7, Sections 10.3 and 10.4], Case I is essentially worked out in the proof of [21, Theorem 3.14]; let us state a slightly more general result:

Proposition 3.5 *Assume that A is p–torsion free, p–adically complete, and $\mathrm{Spec}(A/p)$ is reduced and connected. Then*

$$\mathrm{End}(\mathrm{dR}) = \begin{cases} \mathrm{Frob}_q^{\mathbb{N}} & \text{if } A = \mathbb{Z}_q := W(\mathbb{F}_q), \\ \mathrm{id} & \text{otherwise.} \end{cases}$$

In [21, Section 2.3], a Frobenius map is constructed on p–adic derived de Rham cohomology when the base is a p–torsion free δ–ring, and it is semilinear with respect to the Frobenius on the base δ–ring. The Frob_q appearing above is the corresponding power of the Frobenius associated with the base δ–ring $\mathbb{Z}_q$; one checks easily that it is $\mathbb{Z}_q$–linear as desired.

Proof Let us use Perf to denote the full subcategory of those A–algebras which are of the form $A\langle X_h^{1/p^\infty} \mid h \in H \rangle$ where H is a set. The proof of [21, Theorem 3.14] shows that

- restricting our de Rham cohomology functor to Perf, we get an injection of endomorphism monoids,
- the restricted de Rham cohomology functor has endomorphism monoid given by a submonoid in $\mathbb{Z}$,
- an element $n \in \mathbb{Z}$ above is characterized by its effect on $R = A\langle X^{1/p^\infty} \rangle$, which sends $X \mapsto X^{p^n}$, and
- the image of the restriction map is contained in $\mathbb{N} \subset \mathbb{Z}$.

Let us assume that $q = p^n$ is in the image of the restriction map. Let $R = A\langle X^{1/p^\infty} \rangle$. Take any $a \in A$ and let us contemplate the map $R \to R/(X - a)$. The induced map of de Rham cohomology is the natural inclusion $R = A\langle X^{1/p^\infty} \rangle \to D$, where D is the algebra obtained by p–completely adjoining divided powers of $X - a$ to R. Extending the map $X \mapsto X^q$ from R to D is the same as requiring the image of $X - a$ to have divided powers. Since in D/p we have $X^p = a^p$, we see that $X^q - a = a^q - a + p \cdot d$ for some $d \in D$. The condition now becomes that $a^q - a$ admits divided powers, as (p) always admits divided powers. One can use the natural surjection $D \to R/(X - a)$ to see that an element $a' \in A$ admits divided powers if and only if its image in D admits divided powers. Therefore the condition becomes that $a^q - a \in A$ should admit divided powers for all $a \in A$. The above implies that in A/p we have

$(x^q - x)^p = 0$ for all $x \in A/p$, since A/p is assumed to be reduced. This is equivalent to all of its elements satisfying $x^q = x$. Now we use Lemma 3.4 to conclude that A/p is actually a subalgebra of $\mathbb{F}_q$, and hence A must be the Witt ring of a perfect subfield inside $\mathbb{F}_q$. □

Remark 3.6 Our argument excludes the existence of the q–Frobenius if there is an element $a \in A$ such that $a^q - a$ does not admit divided powers. For instance, if A/p has a transcendental element over $\mathbb{F}_p$, then there is no functorial endomorphism except for the identity, as claimed in [21, Remark 3.15(3)]. It remains unclear to us, for instance, if the p–Frobenius can exist when $A = \mathbb{Z}_p[\sqrt{p}]$.

Next we turn to Case II, which concerns the (base-changed) de Rham cohomology theory on algebras over a perfect ring of characteristic p. Once again the quasisyntomic descent approach helps us prove the following statement (see Proposition 4.10):

Proposition 3.7 *Let us either*

(1) *assume A is an $\mathbb{F}_p$–algebra and B is an A–algebra, and consider the cohomology theory $\mathrm{dR} \otimes_A B$; or*

(2) *assume $A = k$ is a perfect $\mathbb{F}_p$–algebra and B is a $W_m(k)$–algebra, and consider the cohomology theory $\mathrm{dR}_{m,1} \otimes_{W_m(k)} B$.*

Then the endomorphism monoid of the cohomology theory is a submonoid of $\mathbb{N}(\mathrm{Spec}(B))$, where $\mathbb{N}$ stands for the constant monoid scheme of natural numbers with 1 corresponding to the Frobenius.

Proof We largely follow the strategy from the proof of [21, Theorem 3.14]. Let us temporarily denote the cohomology theory by $\mathcal{F}$.

Note that in both cases $\mathcal{F}$ defines a quasisyntomic sheaf on qSyn_A. For (1) this is Proposition 3.1,[5] and for (2) this is Proposition 4.1. Therefore we can restrict ourselves to the category of QRSP A–algebras to compute the endomorphism monoid.

Next we reduce to one particular QRSP A–algebra: $R = A[X^{1/p^\infty}]/(X)$. To make the reduction, apply the trick in the proof of [11, Proposition 7.10] or [21, Theorem 3.14] to see that, for any QRSP A–algebra S, there exists an explicit QRSP A–algebra $S' = A[X_i^{1/p^\infty}; i \in I]/(f_j; j \in J)$, where f_j is an ind-regular sequence in $A[X_i^{1/p^\infty}; i \in I]$, together with a surjection $R' \to R$ inducing a surjection of their values of the cohomology theory. Hence, for any functorial endomorphism, its effect on $\mathcal{F}(S)$ is determined by that on $\mathcal{F}(S')$. Finally, for each $j \in J$, there exists a map $R \to S'$ sending X^{ℓ/p^n} to $(f_j^{1/p^n})^\ell$. The image of $\mathcal{F}(R)$ under these maps generates $\mathcal{F}(S')$, therefore the effect of a functorial endomorphism is determined by its effect on $\mathcal{F}(R)$.

Lastly we need to understand the effect of a potential functorial endomorphism f on $D := \mathcal{F}(R) = D_{(x)}(B[x^{1/p^\infty}])$, the divided power envelope of (x) in $B[x^{1/p^\infty}]$. From the last four paragraphs of the proof of [21, Theorem 3.14], we see there is a finite disconnection of $\mathrm{Spec}(B)$ such that on the jth component $f(x^\ell) = x^{\ell \cdot p^{n_j}}$ for some natural number n_j. Arguing componentwise, we may assume

---

[5]Since A is p–torsion we can drop the p–completion of the tensor product.

without loss of generality that $f(x^\ell) = x^{\ell \cdot p^N}$ for some natural number N; we need to show that this extends uniquely (assuming functoriality) to the whole of D. The algebra D admits a natural grading; considering the functoriality given by the map $R \to R \otimes_A A[t^{1/p^\infty}]$ sending x^ℓ to $x^\ell \otimes t^\ell$ shows that f must multiply the degree by p^N. Now we claim that, for every $n \in \mathbb{N}$, the effect of f on the set of degree $< p^{n+1}$ parts of D is determined by the effect of f on the set of degree $< p^n$ parts of D, which will finish the proof. To that end, notice that the degree $< p^{n+1}$ parts are generated by $\gamma_{p^{n+1}}(x)$ and the degree $< p^n$ parts. Finally, we look at the map $A[x^{1/p^\infty}]/(X) \to A[y^{1/p^\infty}, z^{1/p^\infty}]/(y,z)$ given by $x^{\ell/p^i} \mapsto (y^{1/p^i} + z^{1/p^i})^\ell$. By comparing the coefficients of $\gamma_{p^{n+N}}(y) \cdot \gamma_{p^{n+N}(p-1)}(z)$ of the equation obtained from functoriality, one sees that the effect of f on $\gamma_{p^{n+1}}(x)$ is pinned down by its effect on $\gamma_{p^n}(x)$ and $\gamma_{p^n(p-1)}(x)$. $\qquad\square$

To illustrate the last sentence of the above proof, let us take $n = 0$ and see how to pin down the effect of f on $\gamma_p(x)$. The functoriality gives us a commutative diagram

$$
\begin{array}{ccc}
D & \xrightarrow{\ \ f\ \ } & D \\
{\scriptstyle x \mapsto (y+z)}\Big\downarrow & & \Big\downarrow{\scriptstyle x \mapsto (y+z)} \\
D \otimes_B D & \xrightarrow{\ f \otimes f\ } & D \otimes_B D
\end{array}
$$

Tracing through commutativity for the element $\gamma_p(x)$, we get that, if $f(\gamma_p(x)) = c \cdot \gamma_{p^{N+1}}(x)$ then

$$
c \cdot \gamma_{p^{N+1}}(y) + \sum_{1 \le j \le p-1} \frac{1}{j!(p-j)!} y^{p^N} \cdot z^{p^N(p-1)} + c \cdot \gamma_{p^{N+1}}(z) = c \cdot \sum_{i+j=p^{N+1}} \gamma_i(y)\gamma_j(z).
$$

Therefore $y^{p^N} \cdot (z^{p^N(p-1)}/(p-1)!) = c \cdot \gamma_{p^N}(y) \cdot \gamma_{p^N(p-1)}(z)$ in $D \otimes_B D$, which clearly pins down

$$
c = \frac{(p^N)! \cdot (p^N(p-1))!}{(p-1)!}.
$$

Similar to Proposition 3.5, if we make a reducedness assumption on B/p then we can further decide which powers of the Frobenius can appear depending on the size of B/p. In Proposition 4.10, using the stacky approach, we will say precisely which powers of the Frobenius are allowed in terms of the map $k \to B^\flat$ for Proposition 3.7(2); see Remark 4.8.

4 Endomorphisms of de Rham cohomology, II

In this section, we use a stacky approach to calculate endomorphisms of de Rham and crystalline cohomology functors in situations where it seems difficult to use only quasisyntomic descent methods.

4.1 Unwinding equivalence

We fix two integers $n, m \ge 1$ and a perfect algebra k as before. The goal of this section is to study endomorphisms of the functor

$$
\mathrm{dR}_{m,n} \colon \mathrm{ARings}_{W_n(k)} \to \mathrm{CAlg}\big(D(W_m(k))\big).
$$

First, we will formulate this as a moduli problem. Let S be a discrete test $W_m(k)$–algebra. We can define a functor $\mathrm{End}_{m,n}$ by

$$\mathrm{End}_{m,n}(S) := \mathrm{End}(\mathrm{dR}_{m,n} \otimes_{W_m(k)} S).$$

This defines a functor $\mathrm{End}_{m,n}$ from $W_m(k)$–algebras to spaces, which a priori is a prestack. Let us study the base-changed crystalline cohomology theory; similar to Proposition 3.1 we have the following:

Proposition 4.1 *The functor* $\mathrm{dR}_{m,n} \otimes_{W_m(k)} S$, *when restricted to* $\mathrm{qSyn}_{W_n(k)}$, *defines a quasisyntomic sheaf.*

Proof Denote the derived crystalline cohomology functor relative to W by $\mathrm{dR}_{\infty,n}$. Then we have $\mathrm{dR}_{m,n} \otimes_{W_m(k)} S \simeq \mathrm{dR}_{\infty,n} \otimes_{W(k)} S$. Using the previous description and the fact that $W(k)$ is p–torsion free, to check the quasisyntomic sheaf property it suffices to work derived modulo p. Since $(\mathrm{dR}_{\infty,n} \otimes_{W(k)} S)/^L p \simeq \mathrm{dR}_{1,n} \otimes_k (S/^L p)$, we may reduce to the case where $m = n = 1$ and S is a 1–truncated animated k–algebra. The proof of Proposition 3.1 works verbatim in this setting as well. □

Lemma 4.2 *The space of endomorphisms* $\mathrm{End}(\mathrm{dR}_{m,n} \otimes_{W_m(k)} S)$ *is discrete.*

Proof Similar to the proof of Lemma 3.3, since $\mathrm{dR}_{m,n} \otimes_{W_m(k)} S$ defines a quasisyntomic sheaf by Proposition 4.1, the claim follows from the fact that, for a large quasisyntomic $W_n(k)$–algebra R, the value $(\mathrm{dR}_{m,n} \otimes_{W_m(k)} S)(R) = \mathrm{dR}_{m,n}(R) \otimes_{W_m(k)} S$ is a discrete algebra. □

On the other hand, let us consider the stack $\mathbb{A}^{1,\mathrm{dR}}$, which will always be viewed as a $W_n(k)$–algebra stack over $W_m(k)$ in this section. We define the following prestack, capturing the endomorphisms of this stack along with the extra algebra structure:

Notation 4.3 For a test $W_m(k)$–algebra S, let us use $\mathscr{S}_{m,n}(S)$ to denote the space (groupoid) of endomorphisms of the stack $\mathbb{A}^{1,\mathrm{dR}}_{(S,n)} := \mathbb{A}^{1,\mathrm{dR}}_{m,n} \times_{\mathrm{Spec}\, W_m(k)} \mathrm{Spec}\, S$ as a $W_n(k)$–algebra stack over $\mathrm{Spec}\, S$.

Proposition 4.4 (unwinding equivalence) *The unwinding functor induces an isomorphism of prestacks*

$$\mathrm{Un} \colon \mathscr{S}_{m,n} \simeq \mathrm{End}_{m,n}.$$

Proof Unwinding provides a map from the left-hand side to the right-hand side. To show that it is an isomorphism, let us fix a test $W_m(k)$–algebra S. The $W_n(k)$–algebra stack $\mathbb{A}^{1,\mathrm{dR}}_S$ by definition is an object of $\mathrm{Fun}_\times(\mathrm{Poly}^{\mathrm{op}}_{W_n(k)}, \mathrm{Stacks}_S)$. Since $\mathbb{A}^{1,\mathrm{dR}}_S$ is an affine stack (Proposition 2.34) and the category $\mathrm{AffStacks}_S$ is a full subcategory of Stacks_S, which is closed under small limits, we note that $\mathbb{A}^{1,\mathrm{dR}}_S$ is classified by an object of the full subcategory $\mathrm{Fun}_\times(\mathrm{Poly}^{\mathrm{op}}_{W_n(k)}, \mathrm{AffStacks}_S)$. By Remark 2.18, the global section functor induces an equivalence of ∞–categories $\mathrm{AffStacks}_S \simeq \mathrm{coSCR}^{\mathrm{op}}_S$, where the latter denotes the ∞–category of cosimplicial S–algebras. Therefore $\mathbb{A}^{1,\mathrm{dR}}_S$ can be equivalently viewed as an object of $\mathrm{Fun}(\mathrm{Poly}_{W_n(k)}, \mathrm{coSCR}_S)$. Hence endomorphisms of $\mathbb{A}^{1,\mathrm{dR}}_S$ as a $W_n(k)$–algebra stack can be computed as endomorphisms of the classifying object, which we may call G, inside the category $\mathrm{Fun}(\mathrm{Poly}_{W_n(k)}, \mathrm{coSCR}_S)$.

Now we look at the S–valued points of $\mathrm{End}_{m,n}$. By properties of left Kan extensions, this is given by endomorphisms of $\mathrm{dR}_{m,n} \otimes_{W_m(k)} S$ as a functor from $\mathrm{Poly}_{W_n(k)}$ to $\mathrm{CAlg}(D(S))$. We can also left Kan extend along the inclusion $\mathrm{Poly}_{W_n(k)} \to \mathrm{qSyn}_{W_n(k)}$ and equivalently consider endomorphisms of the functor $H \colon \mathrm{qSyn}_{W_n(k)} \to \mathrm{CAlg}(D(S))$. By Proposition 4.1, we see that H is a quasisyntomic sheaf.

A basis for the quasisyntomic topology on $\mathrm{qSyn}_{W_n(k)}$ is given by flat algebras over $W_n(k)$ whose reduction modulo p is a QRSP algebra over k. The category of such algebras will be denoted by $\mathrm{QRSP}_{W_n(k)}$. On such algebras, the functor H takes values in discrete rings. By properties of right Kan extension, we obtain that the functor H has a canonical enrichment as a functor $H \colon \mathrm{qSyn}_{W_n(k)} \to \mathrm{coSCR}_S$ and endomorphisms can also be calculated in the category $\mathrm{Fun}(\mathrm{qSyn}_{W_n(k)}, \mathrm{coSCR}_S)$. By Proposition 2.36, we see that restricting along $\mathrm{Poly}_{W_n(k)} \to \mathrm{qSyn}_{W_n(k)}$ now realizes G as the canonical enrichment of $\mathrm{dR}_{m,n} \otimes_{W_m(k)} S$. By properties of left Kan extension, the endomorphisms of H can also be computed as endomorphisms of G in the category $\mathrm{Fun}(\mathrm{Poly}_{W_n(k)}, \mathrm{coSCR}_S)$, which finishes the proof. $\qquad\square$

Proposition 4.5 *The functor $\mathscr{S}_{m,n} \colon \mathrm{Alg}_{W_m(k)} \to \mathcal{S}$ is an fpqc sheaf. In fact, it is a sheaf of sets.*

Proof This follows from Lemma 4.2 and the fact that $\mathbb{A}^{1,\mathrm{dR}}$ is an fpqc stack. $\qquad\square$

Before we proceed further, let us make the following definition. Let $m \geq 1$ be an arbitrary integer fixed as before. Then $\mathbb{G}_a^{\mathrm{perf}}$ represents an fpqc sheaf of rings on the category of $W_m(k)$–algebras.

Definition 4.6 We define a sheaf $\underline{\mathrm{Frob}}_k \colon \mathrm{Alg}_{W_m(k)} \to \mathrm{Sets}$ to be the subsheaf of $\mathcal{H}om_{k\text{–}\mathrm{Alg}}(\mathbb{G}_a^{\mathrm{perf}}, \mathbb{G}_a^{\mathrm{perf}})$ such that, if B is a $W_m(k)$–algebra, then $\underline{\mathrm{Frob}}_k(B)$ is the set of k–algebra scheme maps $\mathbb{G}_{a,B}^{\mathrm{perf}} \to \mathbb{G}_{a,B}^{\mathrm{perf}}$ which is induced by an algebra map $B[x^{1/p^\infty}] \to B[x^{1/p^\infty}]$ that sends x to $\sum_i b_i x^{p^i}$, where the sum ranges over a finite subset in $\mathbb{Z}_{\geq 0}$. The sheaf $\underline{\mathrm{Frob}}_k$ naturally has the structure of a commutative monoid.

Notation 4.7 For a $W_m(k)$–algebra B, we write the symbol Frob^i to mean an element of $i \in \underline{\mathrm{Frob}}_k(B)$. We also write Frob^{i+j} to denote the composition of Frob^i and Frob^j.

Remark 4.8 We note that $\underline{\mathrm{Frob}}_k$ is a subsheaf of the sheafification of the constant monoid $\underline{\mathbb{N}}$. In fact, they are equal when $k = \mathbb{F}_p$, but this is not always the case. One can compute that, given a $W_m(k)$–algebra B, we have

$$\underline{\mathrm{Frob}}_k(B) = \mathrm{Hom}_k(\mathbb{G}_{a,B^\flat}, \mathbb{G}_{a,B^\flat}).$$

In very concrete terms, the right-hand side above is the set of pairs $(\mathcal{P}, i)$, where $\mathcal{P}$ is a partition $B = \prod_{j \in J} B_j$ and $i = (i_j)$ is a function on $\mathrm{Spec}(B)$, which is constant on each $\mathrm{Spec}(B_j)$ taking values in $\mathbb{N}$, satisfying the condition that the map $W_m(k)^\flat = k \to B_j^\flat$ factors through a subfield of the finite field $\mathbb{F}_{p^{i_j}}$.

Consequently, one finds that when k is a perfect *field*, the sheaf $\underline{\mathrm{Frob}}_k$ is representable by either the constant monoid scheme $\mathbb{N}$ or the singleton $\{0\}$, depending on whether k is finite or not.

Proposition 4.9 *There is an isomorphism of sheaf of monoids over k*

$$\mathscr{S}_{1,1} \simeq \underline{\mathrm{Frob}}_k.$$

Proof   Let k be a perfect ring. Let B be an arbitrary k–algebra. By Remark 4.8, our goal is to show that $\mathrm{End}(\mathrm{dR} \otimes_k^L B)$ is just given by $\mathrm{Hom}_k(\mathbb{G}_{a,B}, \mathbb{G}_{a,B})$, where $\mathbb{G}_{a,B}$ is regarded as a k–algebra scheme over B. For the proof, we will use another k–algebra, which we denote by $\mathbb{G}_{a,B}^{\mathrm{perf}}$. More explicitly, $\mathbb{G}_a^{\mathrm{perf}}$ is represented by the affine scheme $\mathrm{Spec}\, B[x^{1/p^\infty}]$; see Example 2.6. Note that we have a natural injection of sets $i : \mathrm{Hom}_k(\mathbb{G}_{a,B}, \mathbb{G}_{a,B}) \to \mathrm{Hom}_k(\mathbb{G}_{a,B}^{\mathrm{perf}}, \mathbb{G}_{a,B}^{\mathrm{perf}})$.

Let us first construct a map $\varphi \colon \mathrm{End}(\mathrm{dR} \otimes_k^L B) \to \mathrm{Hom}_k(\mathbb{G}_{a,B}^{\mathrm{perf}}, \mathbb{G}_{a,B}^{\mathrm{perf}})$. We note that dR restricts to a functor on the full subcategory of k–algebras, which we denote by $\mathrm{Poly}_{\mathrm{perf}}/k$, which consists of perfections of finite-type polynomial algebras over k. If $R \in \mathrm{Poly}_{\mathrm{perf}}/k$, then $\mathrm{dR}_{R/k} \otimes_k B \simeq R \otimes_k B$, which defines a functor from $\mathrm{Poly}_{\mathrm{perf}}/k$ to Alg/B sending R to $R \otimes_k B$. This basically classifies perfect k–algebra ring schemes over $\mathrm{Spec}\, B$, and any endomorphism of $\mathrm{dR} \otimes_k^L B$ induces an endomorphism of this perfect k–algebra ring scheme over $\mathrm{Spec}\, B$, which is just given by $\mathbb{G}_{a,B}^{\mathrm{perf}}$. This constructs the required map φ.

We know that any element in $\mathrm{End}(\mathrm{dR} \otimes_k^L B)$ is uniquely determined by a map f of $\mathbb{A}^{1,\mathrm{dR}}$ as a k–algebra stack over $\mathrm{Spec}\, B$. We also note that there is a natural map $\mathbb{G}_a^{\mathrm{perf}} \to \mathbb{A}^{1,\mathrm{dR}}$ of k–algebra stacks over $\mathrm{Spec}\, B$ (from now on we will omit the B from the subscript to ease our notation). By functoriality of $S \mapsto S_{\mathrm{perf}}$ and the fact that this perfection construction commutes with colimits, it follows that the map f lifts to give a map as below:

$$
\begin{array}{ccc}
\mathbb{G}_a^{\mathrm{perf}} & \xrightarrow{\ \hat{f}\ } & \mathbb{G}_a^{\mathrm{perf}} \\
\downarrow & & \downarrow \\
\mathbb{A}^{1,\mathrm{dR}} & \xrightarrow{\ f\ } & \mathbb{A}^{1,\mathrm{dR}}
\end{array}
$$

Let $u \colon \mathbb{G}_a^{\mathrm{perf}} \to \mathbb{G}_a$ denote the natural map of k–algebra schemes. Then the fiber of the map $\mathbb{G}_a^{\mathrm{perf}} \to \mathbb{A}^{1,\mathrm{dR}}$ identifies with $u^* W[F]$; see [24, Proposition 2.2.6]. Therefore, f is given by a map of the quasi-ideal in $\mathbb{G}_a^{\mathrm{perf}}$ given by $u^* W[F] \to \mathbb{G}_a^{\mathrm{perf}}$, which is of the form of a commutative diagram

$$
\begin{array}{ccc}
u^* W[F] & \longrightarrow & \mathbb{G}_a^{\mathrm{perf}} \\
{\scriptstyle t}\downarrow & & \downarrow{\scriptstyle \varphi(f)} \\
u^* W[F] & \longrightarrow & \mathbb{G}_a^{\mathrm{perf}}
\end{array}
$$

In the above, t is required to be a $\mathbb{G}_a^{\mathrm{perf}}$–module map once the target is given the appropriate $\mathbb{G}_a^{\mathrm{perf}}$–module structure via restricting scalars along $\varphi(f)$. Now inspecting the above diagram at the level of global sections yields that the map φ must factor through i, ie $\varphi(f)$ must be induced by an element of $s \in \mathrm{Hom}_k(\mathbb{G}_a, \mathbb{G}_a)$. From this, it follows that the previous commutative diagram is uniquely determined

by a commutative diagram

$$
\begin{array}{ccc}
W[F] & \longrightarrow & \mathbb{G}_a \\
t' \downarrow & & \downarrow s \\
W[F] & \longrightarrow & \mathbb{G}_a
\end{array}
$$

In the above, t is required to be a $\mathbb{G}_a$–module map once the target is given the appropriate $\mathbb{G}_a$–module structure via restricting scalars along φ. In order to understand the map t', we can therefore apply graded Cartier duality [24, Section 2.4]. We note that $W[F]^* = \mathbb{G}_a$, and thus we get a map of graded group schemes $t'^* \colon \mathbb{G}_a \to \mathbb{G}_a$, where the source group scheme $\mathbb{G}_a$ receives its grading via the $\mathbb{G}_a$–module structure induced by restriction of scalars along s. By easy degree considerations, it follows that there exists a unique $\mathbb{G}_a$–module map t' which fits into the above commutative diagram. Therefore, we obtain the natural bijection $\mathrm{End}(\mathrm{dR} \otimes_k^L B) \simeq \mathrm{Hom}_k(\mathbb{G}_{a,B}, \mathbb{G}_{a,B})$, as desired. $\qquad \square$

Proposition 4.10 *For any $m \geq 1$, there is a natural isomorphism of sheaf of monoids over $W_m(k)$*

$$
\mathscr{S}_{m,1} \simeq \underline{\mathrm{Frob}}_k.
$$

Proof Let B be a $W_m(k)$–algebra. There is a k–algebra scheme over B, which we denote by $\mathbb{G}_{a,B}^{\mathrm{perf}}$, whose underlying affine scheme is $\mathrm{Spec}\, B[x^{1/p^\infty}]$. As in the proof of Proposition 4.9, one also obtains a map $\varphi \colon \mathrm{End}(\mathrm{dR} \otimes_{W_m(k)} B) \to \mathrm{Hom}_k(\mathbb{G}_{a,B}^{\mathrm{perf}}, \mathbb{G}_{a,B}^{\mathrm{perf}})$. It follows from going modulo p and applying Proposition 4.9 that φ actually factors to give a map, again denoted by $\varphi \colon \mathrm{End}(\mathrm{dR} \otimes_{W_m(k)} B) \to \underline{\mathrm{Frob}}_k(B)$. We will argue that this map is a bijection.

By using the stack $\mathbb{A}^{1,\mathrm{dR}}$ and the natural map $\mathbb{G}_a^{\mathrm{perf}} \to \mathbb{A}^{1,\mathrm{dR}}$ in a way similar to the proof of Proposition 4.9, this amounts to the more concrete assertion that there is a unique map t of quasi-ideals in $\mathbb{G}_a^{\mathrm{perf}}$ as in

$$
\begin{array}{ccc}
\mathbb{G}_a^{\mathrm{perf},\sharp} & \longrightarrow & \mathbb{G}_a^{\mathrm{perf}} \\
t \downarrow & & \downarrow \iota(x) \\
\mathbb{G}_a^{\mathrm{perf},\sharp} & \longrightarrow & \mathbb{G}_a^{\mathrm{perf}}
\end{array}
$$

Here $x \in \underline{\mathrm{Frob}}_k(B)$ and $\iota \colon \underline{\mathrm{Frob}}_k(B) \to \mathrm{Hom}_k(\mathbb{G}_{a,B}^{\mathrm{perf}}, \mathbb{G}_{a,B}^{\mathrm{perf}})$ denotes the natural inclusion. Let us write U for the coordinate ring of $\mathbb{G}_a^{\mathrm{perf},\sharp}$. Then U is an $\mathbb{N}[1/p]$–graded Hopf algebra over B. It is also a free algebra over B, where all the homogeneous components are free of rank 1 over B. As a graded B–algebra, U is generated by the basis elements in degree p^i for $i \in \mathbb{Z}$. It is enough to check that, for a fixed $x \in \underline{\mathrm{Frob}}_k(B)$, there exists a unique map t which gives a map of quasi-ideals as above. The existence is clear from definition of $\mathbb{G}_a^{\mathrm{perf},\sharp}$ (Example 2.7) by applying the divided power envelope construction. For the uniqueness, we note that once x is fixed, the above diagram forces the homogeneous elements of degree p^i for $i \leq 0$ to be mapped uniquely. The rest follows from inspecting the comultiplication of U and induction on i (see last paragraph of the proof of Proposition 3.7, as well as the discussion after that proof). $\qquad \square$

Remark 4.11 It is possible to prove Proposition 4.10 by using the methods from [24, Section 3.4], which would essentially amount to proving a similar statement about the quasi-ideal $\mathbb{G}_a^{\mathrm{perf},\sharp} \to \mathbb{G}_a^{\mathrm{perf}}$. It is also possible to reduce to the same statement about quasi-ideals directly from Lemma 2.41 by using the compatibility of the map induced on the animated ring $W(S)/^L p$ via the Frobenius on $W(S)$ with the natural Frobenius operator on any animated k–algebra. This implies that any endomorphism of $\mathbb{A}^{1,\mathrm{dR}}$, as a k–algebra stack, lifts along the map $\mathbb{G}_a^{\mathrm{perf}} \to \mathbb{A}^{1,\mathrm{dR}}$ obtained by taking perfection. This lifting property fails for endomorphisms of $\mathbb{A}^{1,\mathrm{dR}}$ as a $W_2(k)$–algebra stack, leading to extra endomorphisms, as will be constructed in Section 4.2.

4.2 Construction of endomorphisms

This subsection describes the construction of "enough" endomorphisms of de Rham cohomology. Our strategy is to crucially exploit the unwinding equivalence proven in Proposition 4.4 to pass to the world of ring stacks and do a small explicit construction there. We will begin by fixing notation and making some definitions. Since we are interested in endomorphisms, we will ignore the Frobenius twist introduced in Notation 2.28.

Notation 4.12 In this section, we work with a perfect ring k of characteristic $p > 0$. We fix two integers $n, m \geq 1$. We will use W to denote the Witt ring scheme over the fixed base $W_m(k)$. Since m and n are fixed, we will denote $\mathbb{A}^{1,\mathrm{dR}}_{(m,n)}$ simply by $\mathbb{A}^{1,\mathrm{dR}}$ when no confusion is likely to occur.

Definition 4.13 We will let $W[p]$ denote the group scheme underlying the kernel of the multiplication by p map on W.

Definition 4.14 We will let $(1 + W[p])^\times$ denote the monoid scheme underlying $x \in W$ satisfying $px = p$. The multiplication on this monoid scheme is given by simply using the multiplication underlying the ring scheme structure on W.

Proposition 4.15 *Let B be a p–nilpotent ring. Then the monoid scheme $(1 + W[p])^\times$ over $\mathrm{Spec}(B)$ is a group scheme.*

Proof This amounts to saying that, for any ring S with $p^m = 0$ in S for some m, if $x \in W(S)$ satisfies $px = p$ then x must be a unit in the ring $W(S)$. Recall that we have a short exact sequence

$$0 \to W(p \cdot S) \to W(S) \to W(S/p) \to 0,$$

where $W(p \cdot S)$ denotes the Witt ring associated with the ideal (viewed as a nonunital ring) $p \cdot S$. Since $p^m = 0$ in S, the ideal $W(p \cdot S)$ is nilpotent. Therefore it suffices to show the image of x in $W(S/p)$ is a unit, and hence we have reduced to the case where S is of characteristic p. Since $p = V(1)$ in this case, the condition on x reads $V(F(x)) = x \cdot V(1) = V(1)$. Injectivity of V shows that $F(x) = 1$, which implies that x is a unit. $\square$

Construction 4.16 Now we will begin our construction of endomorphisms of $\mathbb{A}^{1,\mathrm{dR}}$ as a $W_n(k)$–algebra stack (over the base $W_m(k)$, which is fixed for this section) when $n \geq 2$. Since Definition 2.32 constructs the above stack as the cone of the quasi-ideal $d : W \xrightarrow{\times p} W$, we will explicitly construct maps at the quasi-ideal level, which can be done purely 1–categorically. We note that there is a natural structure map $(W(k) \xrightarrow{\times p} W(k)) \to (W \xrightarrow{\times p} W)$ of quasi-ideals, which describes the structure of $(W \xrightarrow{\times p} W)$ as a quasi-ideal over k. In the language of quasi-ideals, the natural map $W_n(k) \to k$ can be written as a map $(W(k) \xrightarrow{\times p^n} W(k)) \to (W(k) \xrightarrow{\times p} W(k))$, as described below:

$$
\begin{array}{ccc}
W & \xrightarrow{\times p} & W \\
\uparrow & & \uparrow \\
W(k) & \xrightarrow{\times p} & W(k) \\
{\scriptstyle \times p^{n-1}}\uparrow & & \uparrow \\
W(k) & \xrightarrow{\times p^n} & W(k)
\end{array}
$$

We will construct maps of the quasi-ideal $d : W \xrightarrow{\times p} W$ over the quasi-ideal $W(k) \xrightarrow{\times p^n} W(k)$, as described above. Let F be a homomorphism of the $W(k)$–algebra scheme W. A quasi-ideal map from $d : W \xrightarrow{\times p} W$ to itself can be defined by giving a W–linear map $u : W \to F_* W$ which makes the diagram below commutative:

$$
\begin{array}{ccc}
W & \xrightarrow{\times p} & W \\
{\scriptstyle u}\downarrow & & \downarrow{\scriptstyle F} \\
W & \xrightarrow{\times p} & W
\end{array}
$$

However, we need to make sure that such a map respects the additional structure of being a map of quasi-ideals over $W(k) \xrightarrow{\times p^n} W(k)$, ie the following diagram needs to commute:

$$
\begin{array}{c}
W \xrightarrow{\times p} W \\
{\scriptstyle u}\downarrow \quad\quad \downarrow{\scriptstyle F} \\
W \xrightarrow{\times p} W \\
W(k) \xrightarrow{\times p^n} W(k)
\end{array}
$$

As one checks, for any $n \geq 2$, the only condition this imposes is that $pu(1) = p$. This provides the following map, which we wanted to construct:

$$
(1 + W[p])^\times \cdot F \to \mathscr{S}_{m,n}.
$$

Further, for any $n \geq 2$, the above map is clearly an injection by construction. We point out that it is possible to do such a construction for every $W(k)$–algebra map F of the ring scheme W. Let S be a $W_m(k)$–algebra. Then the element of $\mathscr{S}_{m,n}(B)$ constructed above will be denoted by $u \cdot F$, where u is understood to be an element $u(1) \in (1 + W[p])^\times(B)$. By construction, we see that the composition $(u, F') \circ (v, F)$ is equal to $(uF'(v), F'F)$.

Remark 4.17 In the above picture, if we let $n = 1$ then $u(1)$ is *forced by the diagram* to be equal to 1, and one does not get the extra endomorphisms that were constructed above for $n \geq 2$.

Proposition 4.18 *Let $(1 + W[p])^{\times}$ denote the group scheme as above. There is an injection of (sheaves)* $\coprod_{i \in \underline{\mathrm{Frob}}_k} (1 + W[p])^{\times} \cdot \mathrm{Frob}^i \to \mathrm{End}_{m,n}$ *when $n \geq 2$.*

Proof This follows from [Proposition 4.4](#) and [Construction 4.16](#). □

Remark 4.19 Letting B be a $W_m(k)$–algebra, we construct two natural maps

$$\underline{\mathrm{Frob}}_k(B) \to \mathrm{End}_{W(k)}(W_B^{(1)}) \to \underline{\mathrm{Frob}}_k(B).$$

The first arrow follows from the explicit description given in [Remark 4.8](#), and we simply send powers of the Frobenius to powers of the Frobenius on the Witt ring scheme. To exhibit the second arrow, note that any element in $\mathrm{End}_{W(k)}(W_B^{(1)})$ induces an element in $\mathrm{End}_k([W_B^{(1)}/p]) \simeq \mathrm{End}_k(\mathbb{A}_B^{1,\mathrm{dR}})$, which is equivalent to $\underline{\mathrm{Frob}}_k(B)$ by [Proposition 4.10](#). One easily checks that the composition of the two maps gives the identity on $\underline{\mathrm{Frob}}_k(B)$.

4.3 Calculation of the endomorphism monoid

Throughout this subsection, we will fix k to be a perfect algebra as before. Let $A = W_n(k)$, and let B be a k–algebra. In this subsection, we will show that we have found all the endomorphisms of $\mathrm{dR}_{m,n}$; more precisely, the injection in [Proposition 4.18](#) is an isomorphism.

We need some preparations, starting with understanding the homotopy sheaves associated with $\mathbb{A}^{1,\mathrm{dR}}$. Since $\mathbb{A}^{1,\mathrm{dR}}$ is a 1–stack, we only need to understand π_0 and π_1. Once again, we remind the readers that, since we are interested in endomorphisms, we will ignore the Frobenius twist introduced in [Notation 2.28](#).

Proposition 4.20 *For a test algebra S:*

(1) $\mathbb{A}^{1,\mathrm{dR}}(S) = W(S)/^L p$, *where $W(S)/^L p$ denotes the animated ring obtained by quotienting $W(S)$ by p. We note that the object in the category of animated modules underlying $W(S)/^L p$ can be simply described as $\mathrm{Cofib}(W(S) \xrightarrow{\times p} W(S))$.*

(2) *The sheaf $\pi_1(\mathbb{A}^{1,\mathrm{dR}})$ is representable by $W[p]$, the ideal scheme of p–torsion in the ring scheme W.*

(3) *Over a characteristic-p base, the sheaf $\pi_0(\mathbb{A}^{1,\mathrm{dR}})$ is representable by $\mathbb{G}_a$, where the induced map $W \to \mathbb{G}_a$ is given by the natural projection to the zeroth Witt coordinate.*

Proof (1) By definition, we need to prove that the presheaf $P(S) := W(S)/^L p$ is already an fpqc sheaf of animated rings. It is enough to show that $P(S) := \mathrm{Cofib}(W(S) \xrightarrow{\times p} W(S))$ is a sheaf of animated modules. By noting that $\mathrm{Cofib}(W(S) \xrightarrow{\times p} W(S)) = \mathrm{fib}(W(S)[1] \xrightarrow{\times p} W(S)[1])$, we see that it is enough to prove that the functor $Q(S) := W(S)[1]$ is a sheaf of connective animated modules. For this, we only need to show that $H^1_{\mathrm{fpqc}}(\mathrm{Spec}\, S, W) = 0$.

To this end, note that $W = \varprojlim_n W_n$. By [10, Example 3.1.7 and Proposition 3.1.10] and the fact that F on W is faithfully flat, $W = R\varprojlim_n W_n$. Thus $R\Gamma_{\text{fpqc}}(\text{Spec } S, W) = R\varprojlim_n R\Gamma_{\text{fpqc}}(\text{Spec } S, W_n)$. Now one notes that W_n has a finite filtration with the graded pieces being equal to $\mathbb{G}_a$. Thus

$$R\Gamma_{\text{fpqc}}(\text{Spec } S, W) = R\varprojlim_n R\Gamma_{\text{fpqc}}(\text{Spec } S, W_n) = R\varprojlim_n \Gamma(\text{Spec } S, W_n) = R\varprojlim_n W_n(S)$$

$$= \varprojlim_n W_n(S) = W(S).$$

In particular, $H^1_{\text{fpqc}}(\text{Spec } S, W) = 0$, as desired.

(2) This follows from (1).

(3) In the Witt ring of a characteristic-p ring, $p = VF$. The conclusion follows, since $F : W \to W$ is an fpqc surjection. $\square$

In general, $\pi_0(\mathbb{A}^{1,\text{dR}})$ is given by the sheaf of discrete k–algebras W/p. However, if the base is not of characteristic p, this sheaf stops being representable, as noted below. Nevertheless, Lemma 4.23 will help us extract the necessary information from $\pi_0(\mathbb{A}^{1,\text{dR}})$.

Proposition 4.21 *Let B be a ring such that $p \notin (p^2)$ and let $S = \text{Spec}(B)$. The sheaf*

$$\mathcal{F} := \pi_0(\text{Cone}(\mathbb{G}_{a,S}^\sharp \to \mathbb{G}_{a,S})) \simeq \pi_0(\mathbb{A}^{1,\text{dR}})$$

is not representable by an algebraic space over S.

Proof The isomorphism follows from Proposition 2.38. Since both $\mathbb{G}_a$ and $\mathbb{G}_a^\sharp$ are affine schemes, the hypothetical representing algebraic space would be quasicompact and quasiseparated. Below we show there cannot be such a qcqs algebraic space.

It suffices to prove the statement for B/p^2. Hence we may assume $p^2 = 0$ in B. Since the restriction of our sheaf to B/p–algebras is represented by the affine scheme $\mathbb{G}_{a,B/p}$, using [2, Tag 07V6] we see that the sheaf would in fact be represented by an affine scheme over S. Let us denote its ring of functions by R. The natural map $\mathbb{G}_{a,S} \to \text{Spec}(R)$ induces a map $R \to B[t]$. Reducing the ring map modulo p, we see that the image is $B/p[t^p]$. This implies that an element of the form $t^p + p \cdot g$ must be in the image. On the other hand, we claim that the image of the ring map itself is contained in $\{f \in B[t] \mid f'(t) = 0\}$. Indeed, the two compositions

$$\text{Spec}(B[t, \epsilon]/\epsilon^2) \xrightarrow[t \mapsto t+\epsilon]{t \mapsto t} \text{Spec}(B[t]) \to \mathcal{F}$$

yields the same map as $\epsilon \in B[t, \epsilon]/\epsilon^2$ admits divided powers. This shows that the image of $R \to B[t]$ must be contained in the equalizer of the two maps $B[t] \rightrightarrows B[t, \epsilon]/\epsilon^2$. The identification of this equalizer with those polynomials whose derivative is zero follows from the Taylor expansion. Lastly, to get a contradiction, just observe that, if we let $f = t^p + p \cdot g$, then $f' \neq 0$ as $p \notin (p^2)$; however, we had previously argued that $t^p + p \cdot g$ must be in the image. $\square$

Lemma 4.22 *Let B be a $W(k)$–algebra. We have $W(B)[p] \simeq \mathrm{Hom}_{W(k)}(k, \mathbb{A}^{1,\mathrm{dR}})$, where the right-hand side denotes the space of maps as $W(k)$–algebra stacks over B. Given $\beta \in W(B)[p]$, the corresponding homomorphism of sheaves is modeled by*

$$
\begin{array}{ccc}
W & \xrightarrow{\times p} & W \\
\Big\uparrow{\scriptstyle \times(1+\beta)} & & \Big\uparrow \\
W(k) & \xrightarrow{\times p} & W(k)
\end{array}
$$

Here the constant sheaf of $W(k)$–algebras given by k is viewed as a $W(k)$–algebra stack over B.

Proof Since $\mathbb{A}^{1,\mathrm{dR}}$ is 1–truncated, the right-hand side is classified by $\mathrm{Hom}_k(\mathbb{L}_{k/W(k)}, \pi_1(\mathbb{A}^{1,\mathrm{dR}})[1]) = W(B)[p]$ by Proposition A.6. In this identification we have used Proposition 4.20(2). One checks easily that the maps we constructed in the last sentence exactly correspond to β under the above identification, finishing the proof. $\qquad\square$

Our last preparation is to understand those algebra homomorphisms in $\mathrm{End}_k(\pi_0(\mathbb{A}^{1,\mathrm{dR}}))$ which can be lifted to a $W_n(k)$–algebra homomorphism of $\mathbb{A}^{1,\mathrm{dR}}$. It turns out that liftability as a $W_n(k)$–algebra stack for $n > 1$ *automatically* guarantees liftability as a k–algebra stack, as noted below.

Lemma 4.23 *Let B be a $W_m(k)$–algebra, and let us consider $\mathbb{A}^{1,\mathrm{dR}}$ as a k–algebra stack over B. The two natural maps $\mathrm{End}_{W(k)}(\mathbb{A}^{1,\mathrm{dR}}) \to \mathrm{End}_k(\pi_0(\mathbb{A}^{1,\mathrm{dR}}))$ and $\mathrm{End}_k(\mathbb{A}^{1,\mathrm{dR}}) \to \mathrm{End}_k(\pi_0(\mathbb{A}^{1,\mathrm{dR}}))$ have the same image. In particular, by Proposition 4.10, we know the image is naturally in bijection with the monoid $\underline{\mathrm{Frob}}_k(B)$.*

Proof The image of the first map clearly contains the image of the second. Given $f \in \mathrm{End}_{W(k)}(\mathbb{A}^{1,\mathrm{dR}})$, by composing with the natural map $\iota: k \to \mathbb{A}^{1,\mathrm{dR}}$ we get a natural map $f \circ \iota: k \to \mathbb{A}^{1,\mathrm{dR}}$ of $W(k)$–algebra stacks. In Lemma 4.22, we see that $f \circ \iota$ must be classified by some element $1 + \beta \in 1 + W(B)[p]$. By Proposition 4.15, we can find an inverse $(1+\beta)^{-1} \in (1+W[p])^{\times}$; note that the composition $(1+\beta)^{-1} \circ f \circ \iota$ equals ι. Here we regard an element in $(1 + W[p])^{\times}$ as a $W(k)$–algebra automorphism of $\mathbb{A}^{1,\mathrm{dR}}$ by Construction 4.16. Since these elements in $(1 + W[p])^{\times}$ always induce the identity on π_0, we see that $(1+\beta)^{-1} \circ f$ is a k–algebra automorphism lifting to the same ring homomorphism on $\pi_0(\mathbb{A}^{1,\mathrm{dR}})$ as f. $\quad\square$

Theorem 4.24 *Let $A = W_n(k)$, and suppose that B is a $W_m(k)$–algebra. Then*

$$
\mathrm{End}(\mathrm{dR}_{m,n} \otimes_{W_m(k)} B) = \begin{cases} \coprod_{i \in \underline{\mathrm{Frob}}_k(B)} \mathrm{Frob}^i & \text{if } n = 1, \\ \coprod_{i \in \underline{\mathrm{Frob}}_k(B)} (1 + W[p])^{\times}(B) \cdot \mathrm{Frob}^i & \text{if } n \geq 2. \end{cases}
$$

Here the multiplication law in the second case is given by

$$
(u \cdot \mathrm{Frob}^i) \cdot (v \cdot \mathrm{Frob}^j) = u \cdot \mathrm{Frob}^i(v) \cdot \mathrm{Frob}^{i+j},
$$

where $u, v \in (1 + W[p])^{\times}(B)$.

Remark 4.25 These endomorphism spaces are all discrete, by Lemma 3.3. The above theorem states that the map in Proposition 4.18 is actually an isomorphism. From the above calculation, we also conclude that the sheaf of endomorphism monoids is representable if and only if the sheaf $\underline{\mathrm{Frob}}_k$ is representable. This happens whenever k is a perfect *field*, in which case the representing scheme is a combination of the constant monoid scheme $\mathbb{N}$ and the commutative group scheme $(1 + W[p])^\times$, depending on k and n.

Proof When $n = 1$, this is proved in Proposition 4.10. Below we will assume $n \geq 2$.

Recall that in Proposition 4.4 we have shown that the endomorphisms of our de Rham cohomology functor are the same as the endomorphisms of the $W_n(k)$–algebra stack $\mathbb{A}^{1,\mathrm{dR}}_B$ over $\mathrm{Spec}(B)$. Since the category of $W_n(k)$–algebra stacks is equivalent to the category of sheaves of $W_n(k)$–animated algebras (see Remark 2.14) we will compute the endomorphism of $\mathbb{A}^{1,\mathrm{dR}}_B$ viewed as a sheaf of $W_n(k)$–animated algebras on the fpqc site of $\mathrm{Spec}(B)$.

Composing with the map $\mathbb{A}^{1,\mathrm{dR}} \to \pi_0(\mathbb{A}^{1,\mathrm{dR}})$, we get a natural map

$$\mathrm{Hom}_{W_n(k)}(\mathbb{A}^{1,\mathrm{dR}}, \mathbb{A}^{1,\mathrm{dR}}) \xrightarrow{f_n} \mathrm{Hom}_{W_n(k)}(\mathbb{A}^{1,\mathrm{dR}}, \pi_0(\mathbb{A}^{1,\mathrm{dR}})) = \mathrm{End}_k(\pi_0(\mathbb{A}^{1,\mathrm{dR}})).$$

Here and below, Hom refers to homomorphisms of sheaves respecting the designated structure marked by subscript. By Lemma 4.23, we see that

$$\mathrm{Im}(f_n) = \underline{\mathrm{Frob}}_k(B).$$

We need to understand the fiber of f_n. Take an $i \in \underline{\mathrm{Frob}}_k(B)$; by Proposition A.6, the fiber of f_n over Frob^i is a torsor under

$$\mathrm{Hom}_{\mathbb{A}^{1,\mathrm{dR}}}(\mathbb{L}_{\mathbb{A}^{1,\mathrm{dR}}/W_n(k)}, \pi_1(\mathbb{A}^{1,\mathrm{dR}})[1]).$$

Here the sheaf of $\mathbb{A}^{1,\mathrm{dR}}$–module structure on the sheaf $\pi_1(\mathbb{A}^{1,\mathrm{dR}})$ is via $\mathbb{A}^{1,\mathrm{dR}} \to \pi_0(\mathbb{A}^{1,\mathrm{dR}}) \xrightarrow{\mathrm{Frob}^i} \pi_0(\mathbb{A}^{1,\mathrm{dR}})$. To understand this group, let us utilize the cofiber sequence of cotangent complexes from Proposition A.3 associated with the diagram $W_n(k) \to k \to \mathbb{A}^{1,\mathrm{dR}}$:

$$\mathbb{L}_{k/W_n(k)} \otimes_k \mathbb{A}^{1,\mathrm{dR}} \to \mathbb{L}_{\mathbb{A}^{1,\mathrm{dR}}/W_n(k)} \to \mathbb{L}_{\mathbb{A}^{1,\mathrm{dR}}/k}.$$

By Proposition 4.10, the map $\mathrm{End}_k(\mathbb{A}^{1,\mathrm{dR}}) \to \mathrm{End}_k(\pi_0(\mathbb{A}^{1,\mathrm{dR}}))$ is injective with image $\underline{\mathrm{Frob}}_k(B)$. Therefore, again by Proposition A.6, $\mathrm{Hom}_{\mathbb{A}^{1,\mathrm{dR}}}(\mathbb{L}_{\mathbb{A}^{1,\mathrm{dR}}/k}, \pi_1(\mathbb{A}^{1,\mathrm{dR}})[1]) = 0$, and we get an injection

$$\mathrm{Hom}_{\mathbb{A}^{1,\mathrm{dR}}}(\mathbb{L}_{\mathbb{A}^{1,\mathrm{dR}}/W_n(k)}, \pi_1(\mathbb{A}^{1,\mathrm{dR}})[1]) \hookrightarrow \mathrm{Hom}_{\mathbb{A}^{1,\mathrm{dR}}}(\mathbb{L}_{k/W_n(k)} \otimes_k \mathbb{A}^{1,\mathrm{dR}}, \pi_1(\mathbb{A}^{1,\mathrm{dR}})[1]).$$

The latter is identified with

$$\mathrm{Hom}_k(\mathbb{L}_{k/W_n(k)}, \pi_1(\mathbb{A}^{1,\mathrm{dR}})[1]) = \mathrm{Hom}_k(k[1], \pi_1(\mathbb{A}^{1,\mathrm{dR}})[1]) = \pi_1(\mathbb{A}^{1,\mathrm{dR}})(B) = W[p](B).$$

Here the first identification follows from the fact that $\tau_{\leq 1}\mathbb{L}_{k/W_n(k)} = k[1]$, and the last identification is due to Proposition 4.20(2). Unraveling definitions, for any $u \in (1 + W[p])^\times(B)$, the element $u \cdot \mathrm{Frob}^i$ (see Construction 4.16 and Proposition 4.18) in the fiber of f_n is sent to $u - 1 \in W[p](B)$. One easily

sees that the previous sentence in fact gives a bijection. Therefore the fiber of f_n over Frob^i is exactly $(1 + W[p])^{\times}(B) \cdot \mathrm{Frob}^i$ and finishes the calculation of endomorphism sets.

The multiplication law is checked by chasing through the diagram: on the quasi-ideal model, the homomorphism $u \cdot \mathrm{Frob}^i$ sends an element $x \in W(B)$ to $u \cdot \mathrm{Frob}^i(x)$, and one computes

$$u \cdot \mathrm{Frob}^i(v \cdot F^j(x)) = u \cdot \mathrm{Frob}^i(v) \cdot \mathrm{Frob}^{i+j}(x). \qquad \square$$

Remark 4.26 In the above proof, one does not actually need to work with fpqc sheaves, and the same proof works merely at the level of presheaves. However, if one only wanted to prove Theorem 4.24 in the case when $m = 1$, one could work with fpqc sheaves or quasisyntomic sheaves and use the fact that $\pi_0(\mathbb{A}^{1,\mathrm{dR}}) = \mathbb{G}_a$ (from Proposition 4.20) to simplify the proof and avoid invoking Proposition 4.10 and Lemma 4.23. The case $m = 1$ is sufficient for our application in Section 5.

Corollary 4.27 *Let k be an arbitrary perfect algebra. We consider the functor* $\mathrm{End}_{m,n}$ *from Section 4.1 for a fixed $m \geq 1$. There are natural maps of sheaves* $\mathrm{End}_{m,n'} \to \mathrm{End}_{m,n}$ *for $n' \geq n$, which induces an isomorphism if $n \geq 2$. If $n' > n$, and $n = 1$, then all fibers of this natural map are given by the group scheme $(1 + W[p])^{\times}$. The sheaf $\mathrm{End}_{m,1}$ is $\underline{\mathrm{Frob}}_k$.*

Proof This follows from combining Proposition 4.10 and Theorem 4.24. $\qquad \square$

Remark 4.28 (1) The stabilization of $\mathrm{End}_{m,n}$ for $n \geq 2$ that we see above suggests that lifting to W_n for $n > 2$ gives no extra information on the de Rham cohomology of the special fiber, at least in a functorial sense. In the next section, we will see that the extra information on liftability to second Witt vectors gives a strengthening to Deligne and Illusie's decomposition theorem [12]. Combining these two results, we are led to believe the following dichotomy of possibilities on a follow-up question [12, remarque 2.6(iii)]: either liftability over W_2 always guarantees that the Hodge–de Rham spectral sequence degenerates, or there is a counterexample (necessarily of dimension $\geq p + 1$) which is liftable all the way over W.

(2) If B has characteristic p, then $p = V \circ F$ on $W(B)$. The defining equation $u \cdot p = p$ of $(1 + W[p])^{\times}$ becomes $V(F(u)) = V(1)$. Since V is always injective, the group scheme $(1 + W[p])^{\times}$ over a characteristic-p base becomes $\mathbb{G}_m^{\sharp} := W^{\times}[F]$, namely the Frobenius kernel of the multiplicative group scheme $W^{\times}$.

(3) The above discussion tells us that the functorial automorphism group scheme of the mod p de Rham cohomology theory on $W_2(k)$–algebras is given by $\mathbb{G}_m^{\sharp}$. Note that there is a natural inclusion $\mu_p \to \mathbb{G}_m^{\sharp}$ which induces a product decomposition $\mathbb{G}_m^{\sharp} = \mu_p \times \mathbb{G}_a^{\sharp}$ (see Appendix B). In Theorem 5.4, we will utilize the automorphisms coming from μ_p. The remaining $\mathbb{G}_a^{\sharp}$ worth of automorphisms are related to the Sen operator studied in [5].

Our calculation shows that there is no functorial splitting of the whole mod p derived de Rham complex, as a functor from $W_2(k)$–algebras to $\mathrm{CAlg}(D(k))$, into direct sums of the graded pieces of its conjugate filtrations.

Proposition 4.29 *There is no functorial splitting*

$$\mathrm{dR}_{(-\otimes_{W_2(k)}k)/k} \simeq \bigoplus_{i \in \mathbb{N}_{\geq 0}} \mathrm{Gr}_i^{\mathrm{conj}}(\mathrm{dR}_{(-\otimes_{W_2(k)}k)/k})$$

as a functor from smooth $W_2(k)$–algebras to $\mathrm{CAlg}(D(k))$.

Proof Indeed, if there were such a splitting, we would get an automorphism parametrized by $\mathbb{G}_m$, with the ith graded piece having pure weight i. From the calculation of the endomorphism monoid in Theorem 4.24, this would give us an injection $\mathbb{G}_m \hookrightarrow \mathbb{G}_m^\sharp$. But the Frobenius on $\mathbb{G}_m$ is nonzero, whereas it is zero on $\mathbb{G}_m^\sharp$. Hence we know there is no injective map $\mathbb{G}_m \hookrightarrow \mathbb{G}_m^\sharp$ over any characteristic-p base, getting a contradiction.   □

Remark 4.30 (twisted forms of de Rham cohomology) Theorem 4.24 can be applied to understand a question considered by Antieau and Moulinos on the possible existence of étale twists of the de Rham cohomology functor in some cases: letting k be a perfect ring and B be an ordinary $W_m(k)$–algebra, does there exist a functor $F \colon \mathrm{ARings}_{W_n(k)} \to \mathrm{CAlg}(D(B))$ which is isomorphic to $\mathrm{dR}_{m,n} \otimes_{W_m(k)} B$ étale locally on $\mathrm{Spec}\, B$? We thank Antieau for mentioning this question to us. By Theorem 4.24, such functors are classified by $\mathrm{H}^1_{\text{ét}}(\mathrm{Spec}\, B, (1 + W[p])^\times)$. When $m = 1$ and B is perfect, one can show that $\mathrm{H}^1_{\mathrm{fpqc}}(\mathrm{Spec}\, B, (1 + W[p])^\times) = 0$ by using $(1 + W[p])^\times \simeq \mathbb{G}_m^\sharp \simeq \mu_p \times \mathbb{G}_a^\sharp$ over $\mathrm{Spec}\, B$. So in that case, there does not even exist a nontrivial fpqc twist. However, the cohomology group can be nonzero for some choices of B. It would be interesting to study the corresponding twisted forms of de Rham cohomology which can be seen as new cohomology theories, but that direction is not pursued further in this paper. It would also be interesting to compute $\mathrm{H}^1_{\text{ét}}(\mathrm{Spec}\, B, (1 + W[p])^\times)$ in general for $m > 1$.

5 Application to the Deligne–Illusie decomposition

5.1 Drinfeld's refinement of the Deligne–Illusie decomposition

In this section, we explain how to apply our result from Theorem 4.24 on endomorphisms of the de Rham cohomology functor to recover a recent result of Drinfeld concerning a classical theorem due to Deligne and Illusie [12], and Achinger and Suh [1].

Notation 5.1 Fix a perfect ring k as before, and consider the monoid scheme $\mathrm{End}_{1,n}$ from Corollary 4.27 over k. Let B be a k–algebra and let $\sigma \in \mathrm{End}_{1,n}(B)$. By definition we get an endomorphism induced by σ,

$$\mathrm{dR}_{R/W_n(k)} \otimes_{W_n(k)} B \xrightarrow{\sigma} \mathrm{dR}_{R/W_n(k)} \otimes_{W_n(k)} B,$$

which is functorial in the $W_n(k)$–algebra R.

Definition 5.2 For any $W_n(k)$–algebra R, we define the conjugate filtration $\mathrm{Fil}_i^{\mathrm{conj}}$ on $\mathrm{dR}_{R/W_n(k)} \otimes_{W_n(k)} k$ to be the left Kan extension of the canonical filtration on polynomial (or smooth) $W_n(k)$–algebras.

Lemma 5.3 *Assume* $k \to B$ *is flat. Then* σ *preserves* $\mathrm{Fil}_i^{\mathrm{conj}} \otimes_k B$ *for all* i.

Proof Any morphism must preserve the canonical filtration. If R is a polynomial (or smooth) $W_n(k)$–algebra, one easily shows that the canonical filtration on $\mathrm{dR}_{R/W_n(k)} \otimes_{W_n(k)} B$ is just $\mathrm{Fil}_i^{\mathrm{conj}} \otimes_k B$. □

By Theorem 4.24 and Remark 4.28(4), we have an inclusion of k–schemes $(\mathbb{G}_m^\sharp) \subset \mathrm{End}_{1,2}$. Let $B = \Gamma(\mathbb{G}_m^\sharp, \mathcal{O})$. Then the identity map defines an element $\sigma \in \mathbb{G}_m^\sharp(B)$, which can be regarded as the universal point. By the above discussion, the universal point σ gives rise to a comodule structure on $\mathrm{dR}_{R/W_2(k)} \otimes_{W_2(k)} k$ over the Hopf algebra B, functorial in the $W_2(k)$–algebra R, and the conjugate filtration is an increasing filtration of subcomodules. Alternatively, we may view this as an action of $\mathbb{G}_m^\sharp$ on the mod p de Rham cohomology $\mathrm{dR}_{-/W_2(k)} \otimes_{W_2(k)} k$. We may ask what the effect of the $\mathbb{G}_m^\sharp$–action is on each graded piece of the conjugate filtration, viewed as a functor from the category of $W_2(k)$–algebras to the derived ∞–category of B–comodules. The latter can be defined as the derived ∞–category of quasicoherent sheaves on $B\mathbb{G}_m^\sharp$.

Recall that the category of μ_p–representations is semisimple, with simple objects given by $\mathbb{Z}/p$–worth of powers of the universal character. We follow the convention that the universal character $\mu_p \hookrightarrow \mathbb{G}_m$ has weight 1. The following result was first observed by Drinfeld via prismatization, and communicated to us by Bhatt.

Theorem 5.4 *The action of* $\mathbb{G}_m^\sharp$ *on the* i th *graded piece of the conjugate filtration factors through the natural projection* $\mathbb{G}_m^\sharp \to \mu_p$, *and the resulting* μ_p*–action is of pure weight* $i \in \mathbb{Z}/p$.

This fact also appears in [5, Example 4.7.17], where it is proved using Sen operators. Below we give a different argument:

Proof The derived Cartier isomorphism [3, Proposition 3.5] reduces the proof to showing the statement for $i = 0$ and 1. Since the conjugate filtration is defined via left Kan extension from its values on polynomial algebras, using the classical Cartier isomorphism and Künneth formula, we need only understand the behavior of σ on the cohomology of

$$\mathrm{dR}_{W_2(k)[x]/W_2(k)} \otimes_{W_2(k)} k \simeq \mathrm{dR}_{k[x]/k}.$$

Observe that the whole situation is base changed from $k = \mathbb{F}_p$; we immediately reduce to $k = \mathbb{F}_p$.

According to Construction 2.25, the action of σ is defined via the identification

$$\mathrm{dR}_{\mathbb{Z}/p^2[x]/(\mathbb{Z}/p^2)} \otimes_{\mathbb{Z}/p^2} B \simeq \mathrm{R}\Gamma(\mathbb{A}_B^{1,\mathrm{dR}}, \mathcal{O})$$

and the homomorphism of the $\mathbb{Z}/p^2$–algebra stack over $\mathbb{G}_m^\sharp$ given by the diagram

$$
\begin{array}{ccc}
W_B & \xrightarrow{\times p} & W_B \\
{\scriptstyle \times\sigma}\downarrow & & \downarrow{\scriptstyle \mathrm{id}} \\
W_B & \xrightarrow{\times p} & W_B
\end{array}
$$

Here W_B denotes the Witt ring scheme over the base scheme $\mathbb{G}_m^\sharp$. The first cohomology of $\mathrm{dR}_{\mathbb{F}_p[x]/\mathbb{F}_p}$ is a free rank-1 module over its zeroth cohomology. Therefore all we need to do is

(1) show that the induced map on $\mathrm{H}^0(\mathbb{A}_B^{1,\mathrm{dR}}, \mathcal{O})$ is trivial,

(2) exhibit a nonzero element $v \in \mathrm{H}^1(\mathbb{A}_{\mathbb{F}_p}^{1,\mathrm{dR}}, \mathcal{O})$ which pulls back to a weight-1 element in $\mathrm{H}^1(\mathbb{A}_B^{1,\mathrm{dR}}, \mathcal{O})$.

To avoid confusion, let us define the ring scheme $W := \mathrm{Spec}(\mathbb{F}_p[X_0, X_1, \ldots])$ and the quasi-ideal $W := \mathrm{Spec}(\mathbb{F}_p[Y_0, Y_1, \ldots])$. Here the X_i (and similarly the Y_i) are the Witt coordinates. One easily checks the effect of id^* and $(\times \sigma)^*$ on the elements $X_i \mapsto X_i$ and $Y_0 \mapsto t_0 \cdot Y_0$. Here t_0 denotes the element in B corresponding to the natural projection $\mathbb{G}_m^\sharp \to \mu_p$.

Now (1) is easily verified: $\mathrm{H}^0(\mathbb{A}_B^{1,\mathrm{dR}}, \mathcal{O}) \subset B[X_0, X_1, \ldots]$, and hence invariant under the σ–action.

As for (2), we claim that $1 \otimes Y_0 \in \mathbb{F}_p[X_i, Y_j]$ is a nonzero class in $\mathrm{H}^1(\mathbb{A}_{\mathbb{F}_p}^{1,\mathrm{dR}}, \mathcal{O})$. Here we are using the Čech nerve of $\mathrm{Spec}(\mathbb{F}_p[X_0, X_1, \ldots]) \to \mathbb{A}_{\mathbb{F}_p}^{1,\mathrm{dR}}$ to calculate the cohomology of $\mathbb{A}_{\mathbb{F}_p}^{1,\mathrm{dR}}$; implicitly we have used the fact that the [1]–term of the Čech nerve is given by $\mathrm{Spec}(\mathbb{F}_p[X_i, Y_j])$. Granting this claim, the action of σ sends Y_0 to $t_0 \cdot Y_0$, and hence the action on the class $1 \otimes Y_0$ is via the natural projection $\mathbb{G}_m^\sharp \to \mu_p$ and has weight 1. To prove the claim, we use the maps

$$W[F] = \mathbb{G}_a^\sharp \to W = \mathrm{Spec}(\mathbb{F}_p[Y_0, Y_1, \ldots]) \to \mathbb{G}_a = \mathrm{Spec}(\mathbb{F}_p[Y_0]),$$

where the middle W is a copy of quasi-ideal W. The above maps induce a sequence of abelian group stacks:

$$B\mathbb{G}_a^\sharp \to \mathbb{A}^{1,\mathrm{dR}} \to B\mathbb{G}_a.$$

Recall that there is a canonical identification $\mathrm{H}^1(BG, \mathcal{O}) \simeq \mathrm{Hom}(G, \mathbb{G}_a)$ for affine group schemes G via faithfully flat descent along $* \to BG$. The identity map on $\mathbb{G}_a = \mathrm{Spec}(\mathbb{F}_p[Y_0])$ pulls back to $1 \otimes Y_0 \in \mathbb{F}_p[X_i, X_j]$, which checks that $1 \otimes Y_0$ is a cocycle. Furthermore, recall that the induced map $\mathbb{G}_a^\sharp \to \mathbb{G}_a$ realizes the former as the divided power envelope of the origin inside the latter. In particular it is a nonzero map. From the above identification, this tells us that $1 \otimes Y_0$ pulls back to a nonzero class in $\mathrm{H}^1(B\mathbb{G}_a^\sharp, \mathcal{O})$. Therefore the class $1 \otimes Y_0$ is a nonzero class in $\mathrm{H}^1(\mathbb{A}_{\mathbb{F}_p}^{1,\mathrm{dR}}, \mathcal{O})$.   □

Remark 5.5 Let us mention another way to obtain the above result concerning the $\mathbb{G}_m^\sharp$–action on $\mathrm{dR}_{k[x]/k}$. As explained, the action arises from the action of $\mathbb{G}_m^\sharp$ on $\mathbb{A}^{1,\mathrm{dR}}$ in characteristic p. One can show that the stack underlying $\mathbb{A}_{\mathbb{F}_p}^{1,\mathrm{dR}}$ (without the ring stack structure) decomposes as $\mathbb{G}_a \times B\mathbb{G}_a^\sharp$ (see [6, Proposition 5.12]), and the action of $\mathbb{G}_m^\sharp$ is trivial on $\mathbb{G}_a$ and weight 1 on $B\mathbb{G}_a^\sharp$. This gives the desired statement. We thank the referee for pointing this out to us.

Note that the natural projection $\mathbb{G}_m^\sharp \to \mu_p$ admits a splitting: the Teichmüller lift defines a map of group schemes $\mathbb{G}_m \hookrightarrow W^\times$, which induces a map of group schemes $\mu_p \hookrightarrow \mathbb{G}_m^\sharp$.

Let X be a smooth scheme over $W_2(k)$ and consider the de Rham cohomology of its special fiber (relative to k), which by the above discussion admits a μ_p–action. Now look at the canonical truncation in a range of width at most p. The weights that show up in $\mathbb{Z}/p$ are pairwise distinct, and hence we get a splitting

of the induced conjugate filtration. Therefore the above theorem implies the following improvement of a result due to Achinger and Suh [1, Theorem 1.1], which in turn is a strengthening of Deligne and Illusie's result [12, corollaire 2.4].

Corollary 5.6 (Drinfeld) *Let k be a perfect ring of characteristic $p > 0$, let X be a smooth scheme over $W_2(k)$, and let $a \leq b \leq a + p - 1$. Then the canonical truncation $\tau_{[a,b]}(\Omega^\bullet_{X_k/k})$ splits.*

Note that when $p > 2$, in Achinger and Suh's statement in loc. cit. they need $b < a + p - 1$, so their allowed width needs to be at most $p - 1$. In fact, more generally, we have the following decomposition as a consequence of the $\mathbb{G}_m^\sharp$–action in described in Theorem 5.4.

Corollary 5.7 (Drinfeld; see [1, Remark A.5]) *Let X be a smooth scheme over $W_2(k)$ with special fiber X_k. Then there exists a splitting, functorial in X, in the derived ∞–category of Zariski sheaves on X'_k,*

$$F_{X_k/k,*}(\mathrm{dR}_{X_k/k}) \simeq \bigoplus_{i \in \mathbb{Z}/p} F_{X_k/k,*}(\mathrm{dR}^{\mathrm{weight}=i}_{X_k/k}).$$

Moreover $\mathcal{H}^j(F_{X_k/k,*}(\mathrm{dR}^{\mathrm{weight}=i}_{X_k/k})) \neq 0$ implies $j \equiv i$ in $\mathbb{Z}/p$. Here X'_k is the Frobenius twist of X_k and $F_{X_k/k}$ is the relative Frobenius. In particular, the conjugate spectral sequence of liftable smooth varieties can have nonzero differentials only on the $(mp+1)^{\mathrm{st}}$ pages, where $m \in \mathbb{Z}_{>0}$.

Remark 5.8 Drinfeld observed the results in this subsection by using the "stacky approach" to prismatic crystals (which he calls "prismatization"), which was independently developed by Bhatt and Lurie [5]. Using the prismatization functor, Drinfeld produced an action of μ_p on the de Rham complex of a smooth scheme over k that lifts to $W_2(k)$. Our paper partly grew out of an attempt at making sense of and reproving Drinfeld's theorem without introducing prismatization and taking a very algebraic/categorical approach instead. In [5], this action is obtained in a more geometric way by understanding the prismatization of $\mathrm{Spec}(W_2(k))$.

5.2 Uniqueness of functorial splittings

Corollary 5.6 provides a functorial splitting of the $(p-1)^{\mathrm{st}}$ conjugate filtration of the mod p derived de Rham cohomology of any $W_2(k)$–algebra. On the other hand, the classical Deligne–Illusie splitting also has an ∞–categorical functorial enhancement [20, Theorem 1.3.21 and Proposition 1.3.22], which, in spirit, is more related to the work of Fontaine and Messing [16] and Kato [18].

It is a natural question to ask whether these two splittings agree in a functorial way. By the definitions of these two splittings, we see immediately that they are both compatible with the module structure over the zeroth conjugate filtration, and induced from the splitting of the first conjugate filtration by an averaging process; see the step (a) in proof of [12, théorème 2.1].

Below we will prove that there is a unique way to functorially split the first conjugate filtration, and hence the above two functorial splittings must be the same. To that end, let us fix some notation:

Notation 5.9 Consider the stable ∞–category $\mathrm{Fun}(\mathrm{Alg}^{\mathrm{sm}}_{W_2(k)}, D(k))$, where $\mathrm{Alg}^{\mathrm{sm}}_{W_2(k)}$ is the category of smooth $W_2(k)$–algebras and $D(k)$ is the derived (stable) ∞–category of k–vector spaces. Denote by $\mathcal{O}$ the functor that sends any $W_2(k)$–algebra R to the zeroth conjugate filtration of $\mathrm{dR}_{(R \otimes_{W_2(k)} k)/k}$, which has the structure of a commutative algebra object in $\mathrm{Fun}(\mathrm{Alg}^{\mathrm{sm}}_{W_2(k)}, D(k))$. The functor obtained by considering the first piece of the conjugate filtration will be denoted by M, and viewed as an $\mathcal{O}$–module. We have a natural map $\mathcal{O} \to M$; we denote the cofiber by G, which is the first graded piece of the conjugate filtration, also viewed as an $\mathcal{O}$–module.

Now we have a cofiber sequence of $\mathcal{O}$–modules $\mathcal{O} \to M \to G$.

Theorem 5.10 *In the above notation, there is a unique functorial $\mathcal{O}$–module splitting*

$$M = \mathcal{O} \oplus G$$

in $\mathrm{Fun}(\mathrm{Alg}^{\mathrm{sm}}_{W_2(k)}, D(k))$. *In particular, the splitting of* $\mathrm{Fil}^{\mathrm{conj}}_{p-1}(\mathrm{dR}_{(- \otimes_{W_2(k)} k)/k})$ *obtained in* Corollary 5.6 *and* [20, Theorem 1.3.21] *agree.*

Proof The existence part is provided by either Corollary 5.6 or [20, Theorem 1.3.21]. We focus on the uniqueness part in this proof.

Firstly, we note that it suffices to show the uniqueness of the splitting as a quasisyntomic sheaf on the quasisyntomic site of $W_2(k)$. This is because they are left Kan extended from the polynomial case, and polynomial algebras are quasisyntomic. The site $\mathrm{qSyn}_{W_2(k)}$ admits a basis of large quasisyntomic $W_2(k)$–algebras, so we may restrict our functors to this subclass of $W_2(k)$–algebras and show uniqueness of splitting there. All three functors have discrete value on this subclass of $W_2(k)$–algebras, so $\mathcal{O}$ is a sheaf of ordinary k–algebras given by $R \mapsto R/p$ (up to a Frobenius twist), and M and G are sheaves of ordinary $\mathcal{O}$–modules. We will show that there exists a unique section to the surjection of sheaves of $\mathcal{O}$–modules $M \twoheadrightarrow G$.

Step 1 Consider the algebra $R = W_2(k)[x^{1/p^\infty}]/(x)$. In this case

$$D := \mathrm{dR}_{(R \otimes_{W_2(k)} k)/k} \simeq D_{(x)}(k[x^{1/p^\infty}])$$

is the divided power envelope of (x) in $k[x^{1/p^\infty}]$. This algebra admits a natural grading by the monoid $\mathbb{N}[1/p]$. The values of our sheaves evaluated at R are $\mathcal{O} = k[x^{1/p^\infty}]/(x^p)$, and M is the degree-$[0, 2p)$ part of $D_{(x)}(k[x^{1/p^\infty}])$, whereas G is the degree-$[p, 2p)$ part. One checks easily that G is generated by $\overline{\gamma_p(x)}$ (mod the degree-$[0, p)$ part) as an $\mathcal{O}$–module in this case. We claim that the section necessarily sends this generator to $\gamma_p(x) \in M$. Say the section sends this generator to some element $f(x) \in M$. We look at the two maps of $W_2(k)$–algebras from R to $R \otimes_{W_2(k)} W_2(k)[t^{1/p^\infty}]$ given by $x^m \mapsto x^m \cdot t^m$ and $x^m \mapsto x^m$. The associated mod p derived de Rham cohomology is given by $D \otimes_k k[t^{1/p^\infty}]$. Since the corresponding maps of values of G are

$$\overline{\gamma_p(x)} \mapsto \overline{\gamma_p(tx)} = t^p \overline{\gamma_p(x)} \quad \text{and} \quad \overline{\gamma_p(x)} \mapsto \overline{\gamma_p(x)},$$

functoriality tells us that $t^p f(x) = f(tx) \in D \otimes_k k[t^{1/p^\infty}]$. This implies that $f(x)$ is a homogeneous degree-p element in M which maps to $\overline{\gamma_p(x)} \in G$. Therefore it must be $\gamma_p(x) \in M$.

Step 2 Next we consider the algebra $R_n = W_2(k)[x_i^{1/p^\infty}; i = 1, \ldots, n]/(\sum_{i=1}^n x_i)$. In this case, define

$$D_n := \mathrm{dR}_{(R_n \otimes_{W_2(k)} k)/k} = D_{(\sum_{i=1}^n x_i)}(k[x_i^{1/p^\infty}]).$$

Then the value of our sheaves evaluated at R_n is given by $\mathcal{O} = k[x_i^{1/p^\infty}]/(\sum_{i=1}^n x_i^p)$, and $M = \mathcal{O} \cdot \{1, \gamma_p(\sum_{i=1}^n x_i)\}$ whereas $G = \mathcal{O} \cdot \overline{\gamma_p(\sum_{i=1}^n x_i)}$. In this case, we claim that the section necessarily sends $\gamma_p(\sum_{i=1}^n x_i)$ to $\sum_{i=1}^n \gamma_p(x_i)$. Note that this sum makes sense as an element in D_n, and in fact is in M. For instance, one may repeatedly use $\gamma_p(x + y) = \sum_{i=0}^p \gamma_i(x) \cdot \gamma_{p-i}(y)$ to see this. Now to show the above claim, we first use the same argument as in the previous paragraph to see that the section of $\overline{\gamma_p(\sum_{i=1}^n x_i)}$ is necessarily a homogeneous degree-p element $f(x_i)$. Then we use the functoriality provided by the map $R_n \to R^{\otimes_{W_2(k)} n} = W_2(k)[x_i^{1/p^\infty}; i = 1, \ldots, n]/(x_i; i = 1, \ldots, n)$ to see that the element $g(x_i) := f(x_i) - \sum_{i=1}^n \gamma_p(x_i)$ is a homogeneous degree-p element in the kernel of the induced map $D_n \to D^{\otimes_k n}$. The degree-p part of the kernel is the k–span of $\{x_i^p\}_{i=1}^n$ modulo $k \cdot \sum_{i=1}^n x_i^p$. Finally, using functoriality with respect to switching variables, we see that $g(x_i)$ must be a permutation-invariant element, and hence necessarily 0 unless $n = p = 2$. Therefore, when $n \geq 3$, the associated section is determined. By functoriality, the section associated with R_3 determines the section associated with R_2. This finishes the proof of our claim above.

Step 3 The universal algebra that we need to consider is $R' = W_2(k)[x^{1/p^\infty}, y^{1/p^\infty}]/(x + py)$. Note that $R'/p = R/p \otimes_k k[y^{1/p^\infty}]$, so the values of relevant sheaves are those in Step 1 tensored over k with $k[y^{1/p^\infty}]$. The generator $\overline{\gamma_p(x)} = \overline{\gamma_p(x + py)}$ of G under a functorial section goes to $\gamma_p(x) + g(x, y)$, where $g(x, y) \in k[x^{1/p^\infty}, y^{1/p^\infty}]/(x^p)$ has degree p by the same argument as in Step 1. We claim that $g(x, y) = y^p/(p - 1)!$. To see this, first observe that

$$x_1 + x_2 = (x_1^{1/p} + x_2^{1/p})^p + p \cdot F(x_1, x_2) \quad \text{in } W_2(k)[x_1^{1/p}, x_2^{1/p}],$$

where we view $F(x_1, x_2) \in k[x^{1/p^\infty}, y^{1/p^\infty}]$ as a degree-1 polynomial. Then we see that there is a map $R' \to R_2$ sending x and y to Teichmüller lifts of $x_1 + x_2$ and $F(x_1, x_2)$. The induced map of corresponding D's sends $\gamma_p(x) + g(x, y)$ to $\gamma_p(x_1 + x_2) + g(x_1 + x_2, F(x_1, x_2))$. On the other hand, the functoriality forces this element to be sent to $\gamma_p(x_1) + \gamma_p(x_2)$ by Step 2. Therefore we get a relation

$$\gamma_p(x_1 + x_2) + g(x_1 + x_2, F(x_1, x_2)) = \gamma_p(x_1) + \gamma_p(x_2).$$

Let $h(x, y) = g(x, y) - y^p/(p - 1)! \in k[x^{1/p^\infty}, y^{1/p^\infty}]/(x^p)$, which also has degree p. Combining relations, $h(x_1 + x_2, F(x_1, x_2)) = 0 \in k[x_1^{1/p^\infty}, x_2^{1/p^\infty}]/(x_1^p + x_2^p)$. Applying the next lemma with $x_1 + x_2 = a$ and $x_2 = b$, we conclude that $h(x, y)$ must be 0.

Step 4 Given any large quasisyntomic $W_2(k)$–algebra S, we can find an algebra S' of the form $W_2(k)[X_i^{1/p^\infty}, Y_j^{1/p^\infty}; i \in I, j \in J]/(Y_j + f_j(X_i); j \in J)$ and a surjection $S' \twoheadrightarrow S$ inducing a surjection

of their values on all the relevant sheaves; see the proof of [11, Proposition 7.10] or [21, Theorem 3.14] for details. The value of G in this case is generated, as an $\mathcal{O}$ module, by $\gamma_p(Y_j + f_j(X_i))$ where $j \in J$. By functoriality, we may reduce to the case where $S = W_2(k)[\underline{X}^{1/p^\infty}, Y^{1/p^\infty}]/(Y + f(\underline{X}))$. In this case G is generated over $\mathcal{O}$ by the element $\overline{\gamma_p(Y + f(\underline{X})e)}$; we want to show the section is forced on this element. Observe that any element in $W_2(k)[\underline{X}^{1/p^\infty}, Y^{1/p^\infty}]$ can be written as $[P_1] + p \cdot [P_2]$, a Teichmüller lift plus p times another Teichmüller lift. Therefore we can define a map $R' \to S$ sending X to $[P_1]$ and Y to $[P_2]$. Then we see that the section of $\overline{\gamma_p(Y + f(\underline{X}))}$ must be $\gamma_p(P_1) + P_2^p/(p-1)!$ by Step 3. This shows the rigidity, as desired. $\qquad\square$

Lemma 5.11 *Suppose that* $F(a, b) \in k[a^{1/p^\infty}, b^{1/p^\infty}]$ *is the degree-1 element such that its lift* $\widetilde{F}$ *to* $W_2(k)[a^{1/p^\infty}, b^{1/p^\infty}]$ *satisfies*

$$(a - b)^p + b^p = a^p + p \cdot \widetilde{F}(a^p, b^p) \quad \text{in } W_2(k)[a^{1/p^\infty}, b^{1/p^\infty}].$$

Let $H(a, b) \in k[a^{1/p^\infty}, b^{1/p^\infty}]$ *be a degree-p element which does not contain the term* a^p. *Suppose* $H(a, F(a, b)) \in k[a^{1/p^\infty}, b^{1/p^\infty}]$ *is divisible by* a^p. *Then* $H(a, b) = 0$.

Proof Observe that $F(a, b) = \sum_{i=1}^{p-1} c_i \cdot a^{i/p} b^{(p-i)/p}$ with $c_i \neq 0$ for each i. The a–degree of $F(a, b)$ is less than 1, therefore the a–degree of $H(a, F(a, b))$ must be smaller than p unless $H(a, F(a, b)) = 0$ (as $H(a, b)$ does not contain an a^p term). The a^p divisibility now forces $H(a, F(a, b)) = 0$. Considering the b–degree of $H(a, F(a, b))$ shows that, in fact, $H(a, b)$ has to be 0 to begin with. $\qquad\square$

In Step 3 one can alternatively argue using the map $R_{p+1} \to R'$ sending x_1 to x and the rest of the p variables to y.

Appendix A Topos-theoretic cotangent complex

The theory of cotangent complexes appears in many places in the literature. For example, it has been discussed in [17] in the context of simplicial ring objects in a 1–topos, and in [22], where an ∞–categorical theory has been discussed for animated ring objects in spaces. However, in the proof of Theorem 4.24, we required a formalism of cotangent complexes in the generality of animated ring objects in an ∞–topos. In this appendix, we will sketch a formalism of cotangent complexes in the above generality and its very basic properties, which is sufficient for the proof of Theorem 4.24. Our exposition basically uses the techniques from [22] and lifts them to the generality we need.

For simplicity, we will focus on the case necessary for our application, where the ∞–topos $\mathcal{X}$ arises as sheaves of spaces on some Grothendieck site $\mathcal{C}$, which will be fixed. As in Definition 2.10, one defines the ∞–category $\mathrm{ARings}(\mathcal{X}) := \mathrm{ARings}(\mathcal{X})_{\mathbb{Z}}$, which is equivalent to the ∞–category of sheaves of animated rings on $\mathcal{C}$. For a fixed animated ring B in $\mathcal{X}$, one can also consider the ∞–category of connective B–modules in $\mathcal{X}$ defined as the category of sheaves on $\mathcal{C}$ (with values in animated abelian groups) of B–modules.

For $n \geq 0$, an object $F \in \mathrm{ARings}(\mathcal{X})$ will be called n–truncated if $F(c)$ is n–truncated (ie $\pi_i(F(c)) = 0$ for all $i > n$) for all $c \in \mathcal{C}$. We let $\tau_{\leq n} \mathrm{ARings}(\mathcal{X}) \to \mathrm{ARings}(\mathcal{X})$ denote the inclusion of the full subcategory of n–truncated objects in $\mathrm{ARings}(\mathcal{X})$. This admits a left adjoint that sends G to $\tau_{\leq n} G$, which is obtained by n–truncating G as a presheaf first and then applying sheafification.

Construction A.1 (the cotangent complex) Let $A \to B$ be a map in $\mathrm{ARings}(\mathcal{X})$. For any connective B–module M, one can form the trivial square-zero extension $B \oplus M$, which is an object of $\mathrm{ARings}(\mathcal{X})_A$. There is a natural projection map $B \oplus M \to B$, which regards $B \oplus M$ as an object of $(\mathrm{ARings}(\mathcal{X})_A)_{/B}$. One can consider the functor $M \to \mathrm{Maps}_{(\mathrm{ARings}(\mathcal{X})_A)_{/B}}(B, B \oplus M)$. By the adjoint functor theorem, this functor is corepresented by a connective B–module, which we will denote by $\mathbb{L}_{B/A}$.

Remark A.2 Let $A \to B$ be a map in $\mathrm{ARings}(\mathcal{X})$. It follows that $\mathbb{L}_{B/A}$, defined as above, is the sheafification of the presheaf on $\mathcal{C}$ with values in animated abelian groups that sends c to $\mathbb{L}_{B(c)/A(c)}$ for $c \in \mathcal{C}$. It naturally inherits the structure of a sheaf of connective B–modules on $\mathcal{C}$.

Proposition A.3 *For a sequence of morphisms $A \to B \to C$ in $\mathrm{ARings}(\mathcal{X})$, we have a cofiber sequence*

$$\mathbb{L}_{B/A} \otimes_B C \to \mathbb{L}_{C/A} \to \mathbb{L}_{C/B}$$

in the ∞–category of connective C–modules.

Proof This follows from Construction A.1. $\qquad\qquad\square$

Remark A.4 Let $C \in \mathrm{ARings}(\mathcal{X})$. Let $U \to V \to W$ be a cofiber sequence in the ∞–category of connective C–modules. For any connective C–module M, we obtain a long exact sequence

$$\cdots \to \pi_1 \mathrm{Maps}(W, M) \to \pi_1 \mathrm{Maps}(V, M) \to \pi_1 \mathrm{Maps}(U, M) \to \pi_0 \mathrm{Maps}(W, M)$$
$$\to \pi_0 \mathrm{Maps}(V, M) \to \pi_0 \mathrm{Maps}(U, M).$$

Definition A.5 (square-zero extensions) Let $A \in \mathrm{ARings}(\mathcal{X})$ and $B \in \mathrm{ARings}(X)_A$. Let M be a connective B–module. A square-zero extension of B by M is classified by $\mathrm{Maps}_B(\mathbb{L}_{B/A}, M[1])$, where the maps are considered in the ∞–category of connective B–modules. By Construction A.1, square-zero extensions can be equivalently classified by $\mathrm{Maps}_{(\mathrm{ARings}(\mathcal{X})_A)_{/B}}(B, B \oplus M[1])$. Given $s \colon B \to B \oplus M[1]$ which gives a section to the projection, the pullback $B' := B \otimes_{B \oplus M[1]} B$ recovers the total space of the square-zero extension, where B maps to $B \oplus M[1]$ via s and the zero section. The fiber of $B' \to B$ can be identified with M with the natural structure of an A–module.

Proposition A.6 *Let $C \in \mathrm{ARings}(\mathcal{X})_A$ and let $B' \to B$ in $\mathrm{ARings}(\mathcal{X})_A$ be a square-zero extension of B by a connective B–module M. There is a natural map $\mathrm{Maps}_{\mathrm{ARings}(\mathcal{X})_A}(C, B') \to \mathrm{Maps}_{\mathrm{ARings}(\mathcal{X})_A}(C, B)$ such that the nonempty fibers are torsors under the group $\mathrm{Maps}_C(\mathbb{L}_{C/A}, M)$, where the maps are taken in the category of connective C–modules. The C–module structure on M is obtained via the map $C \to B$ over which the fiber is being taken.*

Proof Unwrapping the definitions and using the fact that the mapping spaces are ∞–groupoids, one can reduce to checking this in the case when $B' = B \oplus M$ is the trivial square-zero extension. Fix a map $C \to B$. We need to show that $\mathrm{Maps}_{(\mathrm{ARings}(\mathcal{X})_A)/B}(C, B \oplus M)$ is equivalent to $\mathrm{Maps}_C(\mathbb{L}_{C/A}, M)$. For this, we note that pulling back along $C \to B$ gives an equivalence $\mathrm{Maps}_{(\mathrm{ARings}(\mathcal{X})_A)/B}(C, B \oplus M) \simeq \mathrm{Maps}_{(\mathrm{ARings}(\mathcal{X})_A)/C}(C, C \oplus M)$. By definition, $\mathrm{Maps}_{(\mathrm{ARings}(\mathcal{X})_A)/C}(C, C \oplus M) \simeq \mathrm{Maps}_C(\mathbb{L}_{C/A}, M)$, which gives the conclusion. $\qquad\square$

Remark A.7 For any object $A \in \mathrm{ARings}(\mathcal{X})_A$, one can use the truncation functors to build a sequence of square-zero extensions $\cdots \to \tau_{\leq n+1} A \to \tau_{\leq n} A \to \tau_{\leq n-1} A \to \cdots \to \tau_{\leq 0} A = \pi_0(A)$. This can be seen by first showing a similar statement at the presheaf level and then sheafifying; at the presheaf level, the statement follows from the analogous statement for animated rings; see [22, Proposition 3.3.6]. In particular, if $A \in \mathrm{ARings}(\mathcal{X})_A$ is 1–truncated, then A is a square-zero extension of $\pi_0(A)$ by $\pi_1(A)[1]$, where the latter is viewed as a connective $\pi_0(A)$–module in $\mathcal{X}$.

Appendix B A product formula for $(1 + W[p])^\times$ in characteristic $p > 0$

The group scheme $(1 + W[p])^\times$ was defined in Definition 4.14. Working over a fixed base ring of characteristic $p > 0$, this group scheme is isomorphic to $W^\times[F]$; see Remark 4.28(2). The following proposition was stated in [14, Lemma 3.3.4] and a more general proposition over $\mathbb{Z}_p$ has been proven in [15, Proposition B.5.6] by using the logarithm constructed in loc. cit.; see also [5, Lemma 3.5.18]. Let us give a more direct argument in characteristic p that does not use the logarithm and is closer to deformation theory in spirit:

Proposition B.1 *There exists a natural isomorphism $W^\times[F] \simeq W[F] \times \mu_p$ over any base ring of characteristic p.*

Proof Note that given any nonunital ring $(I, +, \cdot)$, one can define a monoid associated to it, which will be denoted by I'. At the level of underlying sets, $I' := I$, but the composition $x * y$ is defined to be $x + y + x \cdot y$. Using the above construction along with the Yoneda lemma produces a functor from the category of nonunital ring schemes (eg ideals in unital ring schemes) to the category of monoid schemes. Note that we have a short exact sequence

$$0 \to W[F] \to W[F] \xrightarrow{f} \alpha_p \to 0$$

of group schemes. Moreover, the map $f : W[F] \to \alpha_p$ is a map of nonunital ring schemes when $W[F]$ and $\alpha_p \simeq \mathbb{G}_a[F]$ are both equipped with their natural nonunital ring scheme structures. Applying the functor we constructed before, we obtain a map $f' : W^\times[F] \to \mu_p$. It is clear that f' is surjective. The map f' can be identified with projection to zeroth Witt coordinate: given any test algebra S and an element $x \in W[F](S)$, f' sends $1 + x$ to $1 + x_0$, where x_0 is the zeroth Witt coordinate. In particular, the map $\mu_p \to W^\times[F]$ given by the Teichmüller lift is a section to f'. It remains to identify $\mathrm{Ker}\, f'$ with $W[F]$ as a group scheme. This follows from the lemma below.

Lemma B.2 *For the map $f : W[F] \to \alpha_p$, the ideal* $\operatorname{Ker} f$ *is a square-zero ideal.*

Proof We note that the multiplication in $W[F]$ is inherited from the ring scheme W. Let S be a test algebra of characteristic p and let $m, n \in (\operatorname{Ker} f)(S)$. Then $m = V(m')$ and $n = V(n')$ for some $m', n' \in W(S)$. Here V denotes the Verschiebung operator. We have $m \cdot n = V(m') \cdot n = V(m' \cdot F(n)) = 0$, since $F(n) = 0$. $\qquad\square$

The proposition now follows, since we obtain a split exact sequence of group schemes

$$0 \to W[F] \to W^{\times}[F] \xrightarrow{f'} \mu_p \to 0. \qquad\square$$

References

[1] **P Achinger**, **J Suh**, *Some refinements of the Deligne–Illusie theorem*, Algebra Number Theory 17 (2023) 465–496 MR Zbl

[2] **P Belmans**, **A J de Jong**, et al., *The Stacks project*, electronic reference (2005–) Available at http://stacks.math.columbia.edu

[3] **B Bhatt**, *p–adic derived de Rham cohomology*, preprint (2012) arXiv 1204.6560

[4] **B Bhatt**, *Prismatic F–gauges*, lecture notes (2022) Available at https://www.math.ias.edu/~bhatt/teaching/mat549f22/lectures.pdf

[5] **B Bhatt**, **J Lurie**, *Absolute prismatic cohomology*, preprint (2022) arXiv 2201.06120

[6] **B Bhatt**, **J Lurie**, *The prismatization of p–adic formal schemes*, preprint (2022) arXiv 2201.06124

[7] **B Bhatt**, **J Lurie**, **A Mathew**, *Revisiting the de Rham–Witt complex*, Astérisque 424, Soc. Math. France, Paris (2021) MR Zbl

[8] **B Bhatt**, **M Morrow**, **P Scholze**, *Integral p–adic Hodge theory*, Publ. Math. Inst. Hautes Études Sci. 128 (2018) 219–397 MR Zbl

[9] **B Bhatt**, **M Morrow**, **P Scholze**, *Topological Hochschild homology and integral p–adic Hodge theory*, Publ. Math. Inst. Hautes Études Sci. 129 (2019) 199–310 MR Zbl

[10] **B Bhatt**, **P Scholze**, *The pro-étale topology for schemes*, from "De la géométrie algébrique aux formes automorphes, I" (J-B Bost, P Boyer, A Genestier, L Lafforgue, S Lysenko, S Morel, B C Ngo, editors), Astérisque 369, Soc. Math. France, Paris (2015) 99–201 MR Zbl

[11] **B Bhatt**, **P Scholze**, *Prisms and prismatic cohomology*, Ann. of Math. 196 (2022) 1135–1275 MR Zbl

[12] **P Deligne**, **L Illusie**, *Relèvements modulo p^2 et décomposition du complexe de de Rham*, Invent. Math. 89 (1987) 247–270 MR Zbl

[13] **V Drinfeld**, *A stacky approach to crystals*, preprint (2018) arXiv 1810.11853

[14] **V Drinfeld**, *Prismatization*, preprint (2020) arXiv 2005.04746

[15] **V Drinfeld**, *A 1–dimensional formal group over the prismatization of* $\operatorname{Spf} \mathbb{Z}_p$, preprint (2021) arXiv 2107.11466

[16] **J-M Fontaine**, **W Messing**, *p–adic periods and p–adic étale cohomology*, from "Current trends in arithmetical algebraic geometry" (K A Ribet, editor), Contemp. Math. 67, Amer. Math. Soc., Providence, RI (1987) 179–207 MR Zbl

[17] **L Illusie**, *Complexe cotangent et déformations, I*, Lecture Notes in Math. 239, Springer (1971) MR Zbl

[18] **K Kato**, *On p–adic vanishing cycles (application of ideas of Fontaine–Messing)*, from "Algebraic geometry" (T Oda, editor), Adv. Stud. Pure Math. 10, North-Holland, Amsterdam (1987) 207–251 MR Zbl

[19] **D Kubrak**, **A Prikhodko**, *p–adic Hodge theory for Artin stacks*, preprint (2021) arXiv 2105.05319

[20] **D Kubrak**, **A Prikhodko**, *Hodge-to–de Rham degeneration for stacks*, Int. Math. Res. Not. 2022 (2022) 12852–12939 MR Zbl

[21] **S Li**, **T Liu**, *Comparison of prismatic cohomology and derived de Rham cohomology*, J. Eur. Math. Soc. (online publication October 2023)

[22] **J Lurie**, *Derived algebraic geometry* (2004) Available at http://people.math.harvard.edu/~lurie/papers/DAG.pdf

[23] **J Lurie**, *Higher topos theory*, Annals of Mathematics Studies 170, Princeton Univ. Press (2009) MR Zbl

[24] **S Mondal**, $\mathbb{G}_a^{\mathrm{perf}}$*–modules and de Rham cohomology*, Adv. Math. 409 (2022) art. id. 108691 MR Zbl

[25] **S Mondal**, *Reconstruction of the stacky approach to de Rham cohomology*, Math. Z. 302 (2022) 687–693 MR Zbl

[26] **T Moulinos**, **M Robalo**, **B Toën**, *A universal Hochschild–Kostant–Rosenberg theorem*, Geom. Topol. 26 (2022) 777–874 MR Zbl

[27] **A Petrov**, *Non-decomposability of the de Rham complex and non-semisimplicity of the Sen operator*, preprint (2023) arXiv 2302.11389

[28] **C Simpson**, *Homotopy over the complex numbers and generalized de Rham cohomology*, from "Moduli of vector bundles" (M Maruyama, editor), Lecture Notes in Pure and Appl. Math. 179, Dekker, New York (1996) 229–263 MR Zbl

[29] **B Toën**, *Champs affines*, Selecta Math. 12 (2006) 39–135 MR Zbl

Morningside Center of Mathematics and Hua Loo-Keng Key Laboratory of Mathematics
Chinese Academy of Sciences
Beijing, China

Department of Mathematics, University of British Columbia
Vancouver, BC, Canada

lishizhang@amss.ac.cn,   smondal@umich.edu

The nonabelian Brill–Noether divisor on $\overline{\mathcal{M}}_{13}$ and the Kodaira dimension of $\overline{\mathcal{R}}_{13}$

GAVRIL FARKAS
DAVID JENSEN
SAM PAYNE

We highlight several novel aspects of the moduli space of curves of genus 13, the first genus g where phenomena related to $K3$ surfaces no longer govern the birational geometry of $\overline{\mathcal{M}}_g$. We compute the class of the nonabelian Brill–Noether divisor on $\overline{\mathcal{M}}_{13}$ of curves that have a stable rank-two vector bundle with canonical determinant and many sections. This provides the first example of an effective divisor on $\overline{\mathcal{M}}_g$ with slope less than $6 + 10/g$. Earlier work on the slope conjecture suggested that such divisors may not exist. The main geometric application of our result is a proof that the Prym moduli space $\overline{\mathcal{R}}_{13}$ is of general type. Among other things, we also prove the Bertram–Feinberg–Mukai and the strong maximal rank conjectures on $\overline{\mathcal{M}}_{13}$.

14H10; 14T20

1 Introduction

One of the defining achievements of modern moduli theory is the result due to Harris, Mumford and Eisenbud [27; 16] that $\overline{\mathcal{M}}_g$ is of general type for $g \geq 24$. An essential step in their proof is the calculation of the class of the *Brill–Noether divisor* $\overline{\mathcal{M}}_{g,r}^d$ consisting of those curves X of genus g such that $G_d^r(X) \neq \varnothing$ in the case $\rho(g, r, d) := g - (r+1)(g-d+r) = -1$. Recall that the *slope* of an effective divisor D on $\overline{\mathcal{M}}_g$ not containing any of the boundary divisors Δ_i in its support is defined as the quantity $s(D) := a / \min_i b_i$,

where $[D] = a\lambda - b_0\delta_0 - \cdots - b_{\lfloor g/2 \rfloor}\delta_{\lfloor g/2 \rfloor} \in CH^1(\overline{\mathcal{M}}_g)$. Eisenbud and Harris [16] showed that the slope of $\overline{\mathcal{M}}_{g,r}^d$ is $a/b_0 = 6 + 12/(g+1)$. After these seminal results from the 1980s, the fundamental question arose whether one can construct effective divisors D on $\overline{\mathcal{M}}_g$ of slope $s(D) < 6 + 12/(g+1)$ by using conditions defined in terms of *higher rank* vector bundles on curves.

Each effective divisor D on $\overline{\mathcal{M}}_g$ of slope $s(D) < 6 + 12/(g+1)$ must contain the locus $\mathcal{K}_g \subseteq \mathcal{M}_g$ of curves lying on a $K3$ surface; see Farkas and Popa [21]. Since curves on $K3$ surfaces possess stable rank-two vector bundles with canonical determinant and unexpectedly many sections (see Lazarsfeld [35], Mukai [38] and Voisin [48]), it is then natural to focus on conditions defined in terms of rank-two vector bundles with canonical determinant.

For a smooth curve X of genus g, let $\mathrm{SU}_X(2, \omega)$ be the moduli space of semistable rank-two vector bundles E on X with $\det E \cong \omega_X$. For $k \geq 0$, Bertram and Feinberg [7, Conjecture, page 2] and Mukai [38, Problem 4.8] conjectured that for a general curve X, the rank-two Brill–Noether locus

$$\mathrm{SU}_X(2, \omega, k) := \{E \in \mathrm{SU}_X(2, \omega_X) : h^0(X, E) \geq k\}$$

has dimension $\beta(2, g, k) := 3g - 3 - \binom{k+1}{2}$. For a general curve X the *Mukai–Petri map*

$$(1) \qquad\qquad \mu_E : \mathrm{Sym}^2 H^0(X, E) \to H^0(X, \mathrm{Sym}^2(E))$$

is injective for each $E \in \mathrm{SU}_X(2, \omega)$; see Teixidor i Bigas [45]. As a consequence, $\mathrm{SU}_X(2, \omega, k)$ has the expected dimension $\beta(2, g, k)$, if it is nonempty. There are numerous partial results on the nonemptiness of $\mathrm{SU}_X(2, \omega, k)$ — see for instance Lange, Newstead and Park [34], Teixidor i Bigas [44] and Zhang [49] — although still no proof in full generality.

Assume now that $3g - 3 = \binom{k+1}{2}$. Then generically, $\mathrm{SU}_X(2, \omega, k)$ consists of finitely many vector bundles, if it is nonempty. We consider the *nonabelian* Brill–Noether divisor $\mathcal{MP}_g$ on $\mathcal{M}_g$ consisting of curves $[X]$ for which there exists $E \in \mathrm{SU}_X(2, \omega_X, k)$ such that the Mukai–Petri map μ_E is not an isomorphism. In this paper, we focus on the first genuinely interesting case,[1]

$$g = 13 \quad \text{and} \quad k = 8.$$

Our first main result proves this case of the Bertram–Feinberg–Mukai conjecture and computes the class of the closure of the nonabelian Brill–Noether divisor.

Theorem 1.1   *A general curve X of genus* 13 *carries exactly three stable vector bundles $E \in \mathrm{SU}_X(2, \omega, 8)$. The closure in $\overline{\mathcal{M}}_{13}$ of the nonabelian Brill–Noether divisor on $\mathcal{M}_{13}$*

$$\mathcal{MP}_{13} := \{[X] \in \mathcal{M}_{13} : \text{there exists an } E \in \mathcal{SU}_X(2, \omega, 8) \text{ with } \mu_E : \mathrm{Sym}^2 H^0(E) \xrightarrow{\not\cong} H^0(\mathrm{Sym}^2(E))\}$$

has slope equal to

$$s([\overline{\mathcal{MP}}_{13}]) = \tfrac{4109}{610} = 6.735\ldots < 6 + \tfrac{10}{13} = 6.769\ldots.$$

---

[1]It is left to the reader to show that in the previous cases $k = 5, 6$, the corresponding divisors $\mathcal{MP}_6$ and $\mathcal{MP}_8$ are supported on the loci, in $\mathcal{M}_6$ and $\mathcal{M}_8$ respectively, of curves failing the Petri theorem.

To explain the significance of this result, we recall that several infinite series of examples of divisors on $\overline{\mathcal{M}}_g$ for $g \geq 10$ with slope less than $6 + 12/(g+1)$ have been constructed in Farkas [17], Farkas and Popa [21], Farkas, Jensen and Payne [19] and Khosla [32], using syzygies on curves. Quite remarkably, the slopes $s(D)$ of all these divisors D on $\overline{\mathcal{M}}_g$ satisfy

$$6 + \frac{10}{g} \leq s(D) < 6 + \frac{12}{g+1}.$$

The slope $6 + 12/(g+1)$ appears as both the slope of the Brill–Noether divisors $\overline{\mathcal{M}}_{g,r}^d$ and as the slope of a Lefschetz pencil of curves of genus g on a $K3$ surface. Similarly, $6 + 10/g$ is the slope of the family of curves $\{X_t'\}_{t \in \mathbb{P}^1}$ in $\Delta_0 \subseteq \overline{\mathcal{M}}_g$ obtained from a Lefschetz pencil $\{X_t\}_{t \in \mathbb{P}^1}$ of curves of genus $g-1$ on a $K3$ surface S by identifying two sections corresponding to basepoints of the pencil. The natural question has been therefore raised in [10, page 2], whether a slight weakening of the Harris–Morrison slope conjecture [26] remains true and the inequality

$$(2) \hspace{4cm} s(D) \geq 6 + \frac{10}{g}$$

holds for every effective divisor D on $\overline{\mathcal{M}}_g$. Results from Farkas and Popa [21] and Tan [43] imply that inequality (2) holds for all $g \leq 12$. In particular, the divisor $\overline{\mathcal{K}}_{10}$ on $\overline{\mathcal{M}}_{10}$ consisting of curves lying on $K3$ surfaces, which was shown in [21] to be the original counterexample to the slope conjecture, satisfies $s(\overline{\mathcal{K}}_{10}) = 7 = 6 + 10/g$. On $\overline{\mathcal{M}}_{12}$, since a general curve of genus 11 lies on a $K3$ surface, it follows that the pencils $\{X_t'\}_{t \in \mathbb{P}^1}$ cover the boundary divisor $\Delta_0 \subseteq \overline{\mathcal{M}}_{12}$, and consequently the inequality (2) holds. Therefore 13 is the smallest genus where inequality (2) can be tested, and Theorem 1.1 provides a negative answer to the question posed in Chen, Farkas and Morrison [10].

1.1 The Kodaira dimension of the Prym moduli space $\overline{\mathcal{R}}_{13}$

One application of Theorem 1.1 concerns the birational geometry of the moduli space $\overline{\mathcal{R}}_g$ of Prym curves of genus g. The Prym moduli space $\mathcal{R}_g$ classifying pairs $[X, \eta]$, where X is a smooth curve of genus g and η is a 2–torsion point in $\mathrm{Pic}^0(X)$, has been classically used to parametrize moduli of abelian varieties via the Prym map $\mathcal{R}_g \to \mathcal{A}_{g-1}$ [6]. The Deligne–Mumford compactification $\overline{\mathcal{R}}_g$ is uniruled for $g \leq 8$ (see Farkas and Verra [23]), and was previously known to be of general type for $g \geq 14$ and $g \neq 16$ (see Bruns [9] and Farkas and Ludwig [20]).[2]

Theorem 1.2 *The Prym moduli space $\overline{\mathcal{R}}_{13}$ is of general type.*

In particular, 13 is the smallest genus g for which it is known that $\overline{\mathcal{R}}_g$ is of general type. The proof of Theorem 1.2 takes full advantage of Theorem 1.1. It also uses the *universal theta divisor* Θ_{13}, defined as

---

[2]The problem of determining the Kodaira dimension of $\overline{\mathcal{R}}_{16}$ remains open. It was proven in Farkas and Ludwig [20] that the Prym–Green conjecture on $\overline{\mathcal{R}}_{16}$ implies that $\overline{\mathcal{R}}_{16}$ is of general type. However, as shown in Chiodo, Eisenbud, Farkas and Schreyer [11, Proposition 4.4], there is strong indication that the Prym–Green conjecture fails in genus 16.

the locus of Prym curves $[X, \eta] \in \mathcal{R}_{13}$ for which there exists a vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$ such that $H^0(X, E \otimes \eta) \neq 0$. In an indirect way (to be explained later), we calculate the class $[\overline{\Theta}_{13}]$ of the closure of Θ_{13} inside $\overline{\mathcal{R}}_{13}$ and show that

(3) $$K_{\overline{\mathcal{R}}_{13}} \in \mathbb{Q}_{>0}\langle \lambda, [\overline{\Theta}_{13}], [\overline{D}_{13:2}], \text{boundary divisors}\rangle,$$

where $D_{13:2}$ is the effective divisor on $\mathcal{R}_{13}$ introduced in Farkas and Ludwig [20] consisting of Prym curves $[X, \eta]$ for which η can be written as the difference of two effective divisors of degree 6 on X. Since λ is big, it follows that $K_{\overline{\mathcal{R}}_{13}}$ is also big. Theorem 1.2 follows, since the singularities of $\overline{\mathcal{R}}_g$ do not impose adjunction conditions [20].

1.2 The strong maximal rank conjecture on $\overline{\mathcal{M}}_{13}$

The proofs of both Theorems 1.1 and 1.2 are indirect and proceed through a study of the failure locus of the strong maximal rank conjecture (see Aprodu and Farkas [3]) on $\overline{\mathcal{M}}_{13}$. For a general curve X of genus 13 the Brill–Noether locus $W_{16}^5(X)$ is one-dimensional, and $W_{16}^6(X) = \varnothing$. Counting dimensions shows that the multiplication map

$$\phi_L \colon \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$$

has at least a one-dimensional kernel, since $h^0(X, L^{\otimes 2}) = 2\deg(L) + 1 - g = 20$. The space of pairs $[X, L]$ such that $\mathrm{Ker}(\phi_L)$ is at least two-dimensional therefore has expected codimension 2 in the parameter space $\mathfrak{G}_{16}^5$ of all such pairs $[X, L]$. Since the fibers of the map $\sigma \colon \mathfrak{G}_{16}^5 \to \mathcal{M}_{13}$ are in general one-dimensional, the pushforward of this locus is expected to be a divisor on $\mathcal{M}_{13}$.

Our next result verifies this case of the strong maximal rank conjecture and computes the class of the closure of the divisorial part of the failure locus. This is essential input for the calculation of the nonabelian Brill–Noether divisor class in Theorem 1.1 and hence for the proof of Theorem 1.2.

Theorem 1.3 *The locus of curves $[X] \in \mathcal{M}_{13}$ carrying a line bundle $L \in W_{16}^5(X)$ such that the multiplication map $\phi_L \colon \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$ is not surjective is a proper subvariety of $\mathcal{M}_{13}$, having a divisorial part $\mathfrak{D}_{13}$, whose closure in $\overline{\mathcal{M}}_{13}$ has slope*

$$s(\overline{\mathfrak{D}}_{13}) = \tfrac{5059}{749} = 6.754\ldots < 6 + \tfrac{10}{13}.$$

The proof of Theorem 1.3 takes full advantage of the techniques we developed in [19] in the course of our work on $\overline{\mathcal{M}}_{22}$ and $\overline{\mathcal{M}}_{23}$. To that end, we split Theorem 1.3 in two parts.

Recall that a curve is *treelike* if its dual graph becomes a tree after deleting all loop edges [16, page 364]. We consider a proper moduli stack of generalized limit linear series $\sigma \colon \widetilde{\mathfrak{G}}_{16}^5 \to \widetilde{\mathfrak{M}}_{13}$, where $\widetilde{\mathfrak{M}}_{13}$ is a suitable moduli stack of treelike curves of genus 13 equal to $\mathfrak{M}_{13} \cup \Delta_0 \cup \Delta_1$ in codimension one; see Section 2 for a precise definition. We then construct a morphism of vector bundles over $\widetilde{\mathfrak{G}}_{16}^5$ globalizing

the multiplication maps ϕ_L considered before. The degeneracy locus $\mathfrak{U}$ of this morphism, due to its determinantal nature, carries a virtual class $[\mathfrak{U}]^{\mathrm{virt}}$ of codimension 2 inside $\widetilde{\mathfrak{S}}^5_{16}$. Set

$$[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}} := \sigma_*([\mathfrak{U}]^{\mathrm{virt}}) \in CH^1(\widetilde{\mathcal{M}}_{13}).$$

Theorem 1.4 *The following relation for the virtual class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$ holds:*

$$[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}} = 3(5059\lambda - 749\delta_0 - 3929\delta_1) \in CH^1(\widetilde{\mathcal{M}}_{13}).$$

That the degeneracy locus $\mathfrak{U}$ does not map onto $\mathcal{M}_{13}$ is a particular case of the strong maximal rank conjecture of [3]. We prove this case, along with a stronger result that guarantees that the virtual class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$ is effective, using tropical geometry. In particular, we use the method of tropical independence on chains of loops, as introduced in Jensen and Payne [30; 31]. Our construction of the required tropical independences is similar to the one used in our proof that $\overline{\mathcal{M}}_{22}$ and $\overline{\mathcal{M}}_{23}$ are of general type, with one important innovation. In [19], we were able to ignore certain loops called lingering loops. Here, this seems impossible; there are too few nonlingering loops. This difficulty shows up already in the simplest combinatorial case, which we call the vertex-avoiding case; for a discussion of how we resolve this difficulty, see Remarks 4.3 and 4.11.

Theorem 1.5 *For a general curve $[X] \in \mathcal{M}_{13}$ the map $\phi_L \colon \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$ is surjective for all $L \in W^5_{16}(X)$. Furthermore, there is no component of the degeneracy locus $\mathfrak{U}$ mapping with positive-dimensional fibers onto a divisor in $\widetilde{\mathcal{M}}_{13}$.*

Theorem 1.5 implies that $\widetilde{\mathfrak{D}}_{13}$, defined as the divisorial part of $\sigma(\mathfrak{U})$, represents the class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$. Together with Theorem 1.4, this completes the proof of Theorem 1.3.

The existence of effective divisors of exceptionally small slope on $\overline{\mathcal{M}}_{13}$ has direct applications to the birational geometry of the moduli space $\overline{\mathcal{M}}_{13,n}$ of n–pointed stable curves of genus 13.

Theorem 1.6 *The moduli space $\overline{\mathcal{M}}_{13,n}$ is of general type for $n \geq 9$.*

This improves on Logan's result [36] that $\overline{\mathcal{M}}_{13,n}$ is of general type for $n \geq 11$. It is known that $\overline{\mathcal{M}}_{13,n}$ is uniruled for $n \leq 4$; see Agostini and Barros [1].

1.3 The divisor $\mathfrak{D}_{13}$ and rank-two Brill–Noether loci

The link between Theorems 1.1 and 1.3 involves a reinterpretation of the divisor $\mathfrak{D}_{13}$ in terms of rank-two Brill–Noether theory. Let $\mathcal{SU}_{13}(2, \omega, 8)$ denote the moduli space of pairs $[X, E]$, where $[X] \in \mathcal{M}_{13}$ and $E \in \mathrm{SU}_X(2, \omega, 8)$. Consider the forgetful map

$$\vartheta \colon \mathcal{SU}_{13}(2, \omega, 8) \to \mathcal{M}_{13}, \quad [X, E] \mapsto [X].$$

We will show that ϑ is a generically finite map of degree 3 (Theorem 6.5) and that $\mathcal{SU}_{13}(2, \omega, 8)$ is unirational (Corollary 6.3). The fact that $\overline{\mathcal{M}}_{13}$ possesses a modular cover ϑ of such small degree is surprising; we do not know of parallels for other moduli spaces $\overline{\mathcal{M}}_g$.

We now fix a pair $[X, E] \in \mathcal{SU}_{13}(2, \omega, 8)$ and consider the determinant map

$$d \colon \textstyle\bigwedge^2 H^0(X, E) \to H^0(X, \omega_X).$$

It turns out that for a general $[X, E]$ as above, E is globally generated and the map d is surjective. In particular, $\mathbb{P}(\mathrm{Ker}(d)) \subseteq \mathbb{P}\big(\bigwedge^2 H^0(X, E)\big) \cong \mathbb{P}^{27}$ is a 14–dimensional linear space. Since $h^0(X, \omega_X) = 2h^0(X, E) - 3$, it follows that the set of pairs $[X, E]$ satisfying the condition

(4) $$\mathbb{P}(\mathrm{Ker}(d)) \cap G(2, H^0(X, E)) \neq \varnothing,$$

the intersection being taken inside $\mathbb{P}\big(\bigwedge^2 H^0(X, E)\big)$, is expected to be a divisor on $\mathcal{SU}_{13}(2, \omega, 8)$, and its image under projection by the generically finite map ϑ is expected to be also a divisor on $\mathcal{M}_{13}$. We refer to this locus as the *resonance divisor* $\mathfrak{Res}_{13}$, inspired by the algebraic definition of the *resonance variety*; see Aprodu, Farkas, Papadima, Raicu and Weyman [4, Definition 2.4].

Theorem 1.7   *The closure of the resonance divisor in $\mathcal{M}_{13}$*

$\mathfrak{Res}_{13} := \{[X] \in \mathcal{M}_{13} :$ *there exists an* $E \in \mathcal{SU}_X(2, \omega, 8)$ *with* $\mathbb{P}(\mathrm{Ker}(d)) \cap G(2, H^0(X, E)) \neq \varnothing\}$

is an effective divisor in $\overline{\mathcal{M}}_{13}$. *One has the following equality of divisors on* $\overline{\mathcal{M}}_{13}$:

$$\overline{\mathfrak{Res}}_{13} = \overline{\mathfrak{D}}_{13} + 3 \cdot \overline{\mathcal{M}}^1_{13,7}.$$

Here, we recall that $\overline{\mathcal{M}}^1_{13,7}$ is the Hurwitz divisor of heptagonal curves on $\overline{\mathcal{M}}_{13}$ whose class is computed in Harris and Mumford [27]. The set-theoretic inclusion $\mathcal{M}^1_{13,7} \subseteq \mathfrak{Res}_{13}$ is relatively straightforward. The multiplicity 3 with which $\overline{\mathcal{M}}^1_{13,7}$ appears in $\mathfrak{Res}_{13}$ is explained by an excess intersection calculation carried out in Section 7, and confirms once more that the degree of the map $\vartheta \colon \mathcal{SU}_{13}(2, \omega, 8) \to \mathcal{M}_{13}$ is 3.

We conclude this introduction by explaining the connection between the resonance divisor $\mathfrak{Res}_{13}$ and Theorems 1.1 and 1.3. On the one hand, using Farkas and Rimányi [22] the class $[\widetilde{\mathfrak{Res}}_{13}]$ of the closure of $\mathfrak{Res}_{13}$ in $\widetilde{\mathcal{M}}_{13}$ can be computed in terms of the generators of $CH^1(\widetilde{\mathcal{M}}_{13})$ and a tautological class $\vartheta_*(\gamma)$, where γ is the pushforward of the second Chern class of the (normalized) universal rank-two vector bundle on the universal curve over a suitable compactification of $\mathcal{SU}_{13}(2, \omega, 8)$; see Definition 7.3 for details. On the other hand, Theorem 1.7 yields an explicit description of $\widetilde{\mathfrak{Res}}_{13}$. By combining this description with Theorem 1.3, we obtain a *second* calculation for the class $[\widetilde{\mathfrak{Res}}_{13}]$. In this way, we indirectly determine the tautological class $\vartheta_*(\gamma)$; see Proposition 7.7. Once the class of $[\widetilde{\mathfrak{Res}}_{13}]$ is known, the calculation of the class of the nonabelian Brill–Noether divisor $[\widetilde{\mathcal{MP}}_{13}]$ (Theorem 1.1) and that of the universal Theta divisor $[\widetilde{\Theta}_{13}]$ on $\overline{\mathcal{R}}_{13}$ (Theorems 1.2 and 8.3) follow from Grothendieck–Riemann–Roch calculations, after checking suitable transversality assumptions.

2 The failure locus of the strong maximal rank conjecture on $\widetilde{\mathcal{M}}_{13}$

We denote by $\overline{\mathfrak{M}}_g$ the moduli stack of stable curves of genus $g \geq 2$ and by $\overline{\mathcal{M}}_g$ the associated coarse moduli space. We work throughout over an algebraically closed field K of characteristic 0 and the Chow groups that we consider are with rational coefficients. Via the isomorphism $CH^*(\overline{\mathfrak{M}}_g) \cong CH^*(\overline{\mathcal{M}}_g)$, we routinely identify cycle classes on $\overline{\mathfrak{M}}_g$ with their pushforward to $\overline{\mathcal{M}}_g$. Recall that for $g \geq 3$ the group $CH^1(\overline{\mathcal{M}}_g)$ is freely generated by the Hodge class λ and by the classes of the boundary divisors $\delta_i = [\Delta_i]$ for $i = 0, \ldots, \lfloor \frac{1}{2}g \rfloor$.

In this section, we realize the virtual divisor class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$ as the pushforward of the virtual class of a codimension 2 determinantal locus inside the moduli space $\widetilde{\mathfrak{G}}^5_{16}$ of limit linear series of type $\mathfrak{g}^5_{16}$ over an open substack $\widetilde{\mathfrak{M}}_{13}$ of $\overline{\mathfrak{M}}_{13}$, which agrees with $\mathfrak{M}_{13} \cup \Delta_0 \cup \Delta_1$ outside a subset of codimension 2. We will use standard terminology from the theory of limit linear series [15], and begin by recalling a few of the basics.

Definition 2.1 Let X be a smooth curve of genus g with $\ell = (L, V) \in G^r_d(X)$ a linear series. The *ramification sequence* of ℓ at a point $q \in X$ is denoted by

$$\alpha^\ell(q) : \alpha^\ell_0(q) \leq \cdots \leq \alpha^\ell_r(q).$$

This is obtained from the *vanishing sequence* $a^\ell(q) : a^\ell_0(q) < \cdots < a^\ell_r(q) \leq d$ of ℓ at q, by setting $\alpha^\ell_i(q) := a^\ell_i(q) - i$ for $i = 0, \ldots, r$. The *ramification weight* of q with respect to ℓ is $\mathrm{wt}^\ell(q) := \sum_{i=0}^r \alpha^\ell_i(q)$. We define $\rho(\ell, q) := \rho(g, r, d) - \mathrm{wt}^\ell(q)$.

A *generalized limit linear series* on a treelike curve X of genus g consists of a collection

$$\ell = \{(L_C, V_C) : C \text{ is a component of } X\},$$

where L_C is a rank-one torsion-free sheaf of degree d on C and $V_C \subseteq H^0(C, L_C)$ is an $(r+1)$–dimensional space of sections satisfying compatibility conditions on the vanishing sequences at the nodes of X; see [16, page 364]. Let $\overline{G}^r_d(X)$ be the variety of generalized limit linear series of type $\mathfrak{g}^r_d$ on X.

In this section we set

(5)
$$g = 13, \quad r = 5, \quad d = 16.$$

Although we are mainly interested in the case $g = 13$, some of the constructions are set up for an arbitrary genus g, making it easier to refer to results from [19].

We denote by $\mathcal{M}_{13,15}^5$ the subvariety of $\mathcal{M}_{13}$ parametrizing curves X such that $W_{15}^5(X) \neq \varnothing$. As explained in [19, Section 3], we have $\mathrm{codim}(\mathcal{M}_{13,5}^5, \mathcal{M}_{13}) \geq 2$.

Let $\Delta_1^\circ \subseteq \Delta_1 \subseteq \overline{\mathcal{M}}_g$ be the locus of curves $[X \cup_y E]$, where X is a smooth curve of genus $g - 1$ and $[E, y] \in \overline{\mathcal{M}}_{1,1}$ is an arbitrary elliptic curve. The point of attachment $y \in X$ is chosen arbitrarily. Furthermore, let $\Delta_0^\circ \subseteq \Delta_0 \subseteq \overline{\mathcal{M}}_g$ be the locus of curves $[X_{yq} := X/y \sim q] \in \Delta_0$, where $[X, q]$ is a smooth curve of genus $g - 1$ and $y \in X$ is an arbitrary point, together with their degenerations $[X \cup_q E_\infty]$, where E_∞ is a rational nodal curve (that is, E_∞ is a nodal elliptic curve and $j(E_\infty) = \infty$). Points of this form comprise the intersection $\Delta_0^\circ \cap \Delta_1^\circ$. We define the following open subset of $\overline{\mathcal{M}}_g$:

$$\overline{\mathcal{M}}_g^\circ := \mathcal{M}_g \cup \Delta_0^\circ \cup \Delta_1^\circ.$$

Along the lines of [19, Section 3], we introduce an even smaller open subspace of $\overline{\mathcal{M}}_g$, over which the calculation of $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$ can be completed. Let $\mathcal{T}_0 \subset \Delta_0^\circ$ be the locus of curves $[X_{yq} := X/y \sim q]$, where either $\overline{G}_d^{r+1}(X) \neq \varnothing$ or $\overline{G}_{d-2}^r(X) \neq \varnothing$. Similarly, let $\mathcal{T}_1 \subseteq \Delta_1^\circ$ be the locus of curves $[X \cup_y E]$, where X is a smooth curve of genus $g - 1$ such that $G_d^{r+1}(X) \neq \varnothing$ or $G_{d-2}^r(X) \neq \varnothing$. We set

$$\widetilde{\mathcal{M}}_g := \overline{\mathcal{M}}_g^\circ \setminus (\overline{\mathcal{M}}_{g,d-1}^r \cup \mathcal{T}_0 \cup \mathcal{T}_1).$$

We define $\widetilde{\Delta}_0 := \widetilde{\mathcal{M}}_g \cap \Delta_0 \subseteq \Delta_0^\circ$ and $\widetilde{\Delta}_1 := \widetilde{\mathcal{M}}_g \cap \Delta_1 \subseteq \Delta_1^\circ$. Note that $\widetilde{\mathcal{M}}_g$ and $\mathcal{M}_g \cup \Delta_0 \cup \Delta_1$ agree away from a set of codimension two in each. We identify $CH^1(\widetilde{\mathcal{M}}_g) \cong \mathbb{Q}\langle \lambda, \delta_0, \delta_1 \rangle$, where λ is the Hodge class, $\delta_0 := [\widetilde{\Delta}_0]$ and $\delta_1 := [\widetilde{\Delta}_1]$.

2.1 Stacks of limit linear series

Let $\widetilde{\mathfrak{G}}_d^r$ be the stack of pairs $[X, \ell]$, where $[X] \in \widetilde{\mathcal{M}}_g$ and ℓ is a (generalized) limit linear series of type g_d^r on the treelike curve X. We consider the proper projection

$$\sigma : \widetilde{\mathfrak{G}}_d^r \to \widetilde{\mathfrak{M}}_g.$$

Over a curve $[X \cup_y E] \in \widetilde{\Delta}_1$, we identify $\sigma^{-1}([X \cup_y E])$ with the variety of (generalized) limit linear series $\ell = (\ell_X, \ell_E) \in \overline{G}_d^r(X \cup_y E)$. The fiber $\sigma^{-1}([X_{yq}])$ over an irreducible curve $[X_{yq}] \in \widetilde{\Delta}_0 \setminus \widetilde{\Delta}_1$ is canonically identified with the variety $\overline{W}_d^r(X_{yq})$ of rank-one torsion-free sheaves L on X_{yq} having degree $d(L) = d$ and $h^0(X_{yq}, L) \geq r + 1$.

Let $\widetilde{\mathfrak{C}}_g \to \widetilde{\mathfrak{M}}_g$ be the universal curve, and let $p_2 : \widetilde{\mathfrak{C}}_g \times_{\widetilde{\mathfrak{M}}_g} \widetilde{\mathfrak{G}}_d^r \to \widetilde{\mathfrak{G}}_d^r$ be the projection map. We denote by $\mathfrak{Z} \subseteq \widetilde{\mathfrak{C}}_g \times_{\widetilde{\mathfrak{M}}_g} \widetilde{\mathfrak{G}}_d^r$ the codimension-two substack consisting of pairs $[X_{yq}, L, z]$, where $[X_{yq}] \in \Delta_0^\circ$, the point z is the node of X_{yq} and $L \in \overline{W}_d^r(X_{yq}) \setminus W_d^r(X_{yq})$ is a *non-locally free* torsion-free sheaf. Let

$$\epsilon : \hat{\mathfrak{C}}_g := \mathrm{Bl}_{\mathfrak{Z}}(\widetilde{\mathfrak{C}}_g \times_{\widetilde{\mathfrak{M}}_g} \widetilde{\mathfrak{G}}_d^r) \to \widetilde{\mathfrak{C}}_g \times_{\widetilde{\mathfrak{M}}_g} \widetilde{\mathfrak{G}}_d^r$$

be the blowup of this locus, and we denote the induced universal curve by

$$\wp := p_2 \circ \epsilon : \widehat{\mathfrak{C}}_g \to \widetilde{\mathfrak{G}}_d^r.$$

The fiber of $\wp$ over a point $[X_{yq}, L] \in \widetilde{\Delta}_0$, where $L \in \overline{W}_d^r(X_{yq}) \setminus W_d^r(X_{yq})$, is the semistable curve $X \cup_{\{y,q\}} R$ of genus g, where R is a smooth rational curve meeting X transversally at y and q.

2.2 A degeneracy locus inside $\widetilde{\mathfrak{G}}_{16}^5$

In order to define the degeneracy locus on $\widetilde{\mathfrak{G}}_{16}^5$ whose pushforward produces $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$, we first choose a Poincaré line bundle $\mathcal{L}$ over the universal curve $\widehat{\mathfrak{C}}_g$ with the following properties:

(i) If $[X \cup_y E] \in \widetilde{\Delta}_1$ and $\ell = (\ell_X, \ell_E) \in \overline{G}_d^r(X \cup E)$ is a limit linear series, then

$$\mathcal{L}_{|[X \cup_y E, \ell]} \in \mathrm{Pic}^d(X) \times \mathrm{Pic}^0(E).$$

(ii) For a point $t = [X_{yq}, L]$, where $[X_{yq}] \in \widetilde{\Delta}_0$ and $L \in \overline{W}_d^r(X_{yq}) \setminus W_d^r(X_{yq})$, thus $L = \nu_*(A)$ for some $A \in W_{d-1}^r(X)$, we have $\mathcal{L}_{|X} \cong A$ and $\mathcal{L}_{|R} \cong \mathcal{O}_R(1)$. Here, $\wp^{-1}(t) = X \cup R$, whereas $\nu \colon X \to X_{yq}$ is the normalization map.

We now introduce two sheaves over $\widetilde{\mathfrak{G}}_d^r$,

$$\mathcal{E} := \wp_*(\mathcal{L}) \quad \text{and} \quad \mathcal{F} := \wp_*(\mathcal{L}^{\otimes 2}).$$

Both $\mathcal{E}$ and $\mathcal{F}$ are locally free; the proof by local analysis in [19, Proposition 3.6] goes through essentially without change.

There is a sheaf morphism over $\widetilde{\mathfrak{G}}_{16}^5$ globalizing the multiplication of sections

(6) $$\phi \colon \mathrm{Sym}^2(\mathcal{E}) \to \mathcal{F}.$$

We denote by $\mathfrak{U} \subseteq \widetilde{\mathfrak{G}}_{16}^5$ the locus where ϕ is not surjective (equivalently, where $\phi^\vee$ is not injective). Due to its determinantal nature, $\mathfrak{U}$ carries a virtual class in the expected codimension 2.

Definition 2.2 We define the virtual divisor class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}} := \sigma_*([\mathfrak{U}]^{\mathrm{virt}})$ as

$$[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}} := \sigma_*(c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)) \in CH^1(\widetilde{\mathfrak{M}}_{13}).$$

If $\mathfrak{U}$ has pure codimension 2, then $\widetilde{\mathfrak{D}}_{13}$ is a divisor on $\widetilde{\mathcal{M}}_{13}$ and $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}} = [\widetilde{\mathfrak{D}}_{13}]$. The following corollary provides a local description of the morphism ϕ.

Corollary 2.3 *The morphism $\phi \colon \mathrm{Sym}^2(\mathcal{E}) \to \mathcal{F}$ has the following description on fibers*:

(i) *For $[X, L] \in \widetilde{\mathfrak{G}}_d^r$, with $[X] \in \mathcal{M}_g \setminus \mathcal{M}_{g,d-1}^r$ smooth, the fibers are*

$$\mathcal{E}_{(X,L)} = H^0(X, L) \quad \text{and} \quad \mathcal{F}_{(X,L)} = H^0(X, L^{\otimes 2}),$$

and $\phi_{(X,L)} \colon \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$ is the usual multiplication map of global sections.

(ii) Suppose $t = (X \cup_y E, \ell_X, \ell_E) \in \sigma^{-1}(\widetilde{\Delta}_1)$, where X is a curve of genus $g - 1$, E is an elliptic curve and $\ell_X = |L_X|$ is the X–aspect of the corresponding limit linear series with $L_X \in W_d^r(X)$ such that $h^0(X, L_X(-2y)) \geq r$. If L_X has no basepoint at y, then

$$\mathcal{E}_t = H^0(X, L_X) \cong H^0(X, L_X(-2y)) \oplus K \cdot u \quad \text{and} \quad \mathcal{F}_t = H^0(X, L_X^{\otimes 2}(-2y)) \oplus K \cdot u^2,$$

where $u \in H^0(X, L_X)$ is any section such that $\mathrm{ord}_y(u) = 0$.

If L_X has a basepoint at y, then

$$\mathcal{E}_t = H^0(X, L_X) \cong H^0(X, L_X(-y)),$$

and the image of $\mathcal{F}_t \to H^0(X, L_X^{\otimes 2})$ is the subspace $H^0(X, L_X^{\otimes 2}(-2y)) \subseteq H^0(X, L_X^{\otimes 2})$.

(iii) Let $t = [X_{yq}, L] \in \sigma^{-1}(\widetilde{\Delta}_0)$ be a point with $q, y \in X$ and let $L \in W_d^r(X_{yq})$ be a locally free sheaf of rank one, such that $h^0(X, \nu^* L(-y - q)) \geq r$, where $\nu \colon X \to X_{yq}$ is the normalization. Then the fibers are described as

$$\mathcal{E}_t = H^0(X, \nu^* L) \quad \text{and} \quad \mathcal{F}_t = H^0(X, \nu^* L^{\otimes 2}(-y - q)) \oplus K \cdot u^2,$$

where $u \in H^0(X, \nu^* L)$ is any section not vanishing at both points y and q.

(iv) Let $t = [X_{yq}, \nu_*(A)]$, where $A \in W_{d-1}^r(X)$, and again set $X \cup_{\{y,q\}} R$ to be the fiber $\wp^{-1}(t)$. Then $\mathcal{E}_t = H^0(X \cup R, \mathcal{L}_{X \cup R}) \cong H^0(X, A)$ and $\mathcal{F}_t = H^0(X \cup R, \mathcal{L}_{X \cup R}^{\otimes 2})$. Furthermore, $\phi(t)$ is the multiplication map on $X \cup R$.

Proof The proof is essentially identical to the proof of [19, Corollary 3.8]; we omit the details. □

2.3 Test curves in $\widetilde{\mathcal{M}}_{13}$

As in [19], the calculation of $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$ is carried out by understanding the restriction of the morphism ϕ along the pullbacks of the three standard test curves F_0, F_{ell} and F_1 inside $\widetilde{\mathcal{M}}_{13}$. Let $[X, q]$ be a general pointed curve of genus $g - 1$ and fix an elliptic curve $[E, y]$. We then define

$$F_0 := \{X_{yq} := X/y \sim q : y \in X\} \subseteq \Delta_0^\circ \subseteq \overline{\mathcal{M}}_g^\circ \quad \text{and} \quad F_1 := \{X \cup_y E : y \in X\} \subseteq \Delta_1^\circ \subseteq \overline{\mathcal{M}}_g^\circ.$$

Furthermore, we define the curve

$$(7) \qquad\qquad F_{\mathrm{ell}} := \{[X \cup_q E_t] : t \in \mathbb{P}^1\} \subseteq \Delta_1 \subseteq \overline{\mathcal{M}}_g,$$

where $\{[E_t, q]\}_{t \in \mathbb{P}^1}$ denotes a pencil of plane cubics and q is a fixed point of the pencil. We record the intersection of these test curves with the generators of $CH^1(\overline{\mathcal{M}}_g)$:

$$F_0 \cdot \lambda = 0, \quad F_0 \cdot \delta_0 = 2 - 2g, \quad F_0 \cdot \delta_1 = 1, \quad F_0 \cdot \delta_j = 0 \quad \text{for } j = 2, \ldots, \lfloor \tfrac{1}{2} g \rfloor,$$

$$F_{\mathrm{ell}} \cdot \lambda = 1, \quad F_{\mathrm{ell}} \cdot \delta_0 = 12, \quad F_{\mathrm{ell}} \cdot \delta_1 = -1, \quad F_{\mathrm{ell}} \cdot \delta_j = 0 \quad \text{for } j = 2, \ldots, \lfloor \tfrac{1}{2} g \rfloor.$$

Note also that $F_1 \cdot \lambda = 0$, $F_1 \cdot \delta_i = 4 - 2g$ and $F_1 \cdot \delta_j = 0$ for $j \neq 1$.

We now describe the pullback $\sigma^*(F_0) \subseteq \widetilde{\mathfrak{G}}^5_{16}$. Having fixed a general pointed curve $[X, q] \in \overline{\mathcal{M}}_{12,1}$, we introduce the variety

$$(8) \qquad Y := \{(y, L) \in X \times W^5_{16}(X) : h^0(X, L(-y-q)) \geq 5\},$$

together with the projection $\pi_1 \colon Y \to X$. Arguing in a way similar to [19, Proposition 3.10], we conclude that Y has pure dimension 2, that is, its actual dimension equals its expected dimension as a degeneracy locus. We consider two curves inside Y, namely

$$\Gamma_1 := \{(y, A(y)) : y \in X, A \in W^5_{15}(X)\} \quad \text{and} \quad \Gamma_2 := \{(y, A(q)) : y \in X, A \in W^5_{15}(X)\},$$

intersecting transversely along finitely many points. We then introduce the variety $\widetilde{Y}$ obtained from Y by identifying for each $(y, A) \in X \times W^5_{15}(X)$, the points $(y, A(y)) \in \Gamma_1$ and $(y, A(q)) \in \Gamma_2$. Let $\vartheta \colon Y \to \widetilde{Y}$ be the projection map.

Proposition 2.4 *With notation as above, there is a birational morphism*

$$f \colon \sigma^*(F_0) \to \widetilde{Y},$$

which is an isomorphism outside $\vartheta(\pi_1^{-1}(q))$. The restriction of f to $f^{-1}(\vartheta(\pi_1^{-1}(q)))$ forgets the aspect of each limit linear series on the elliptic curve E_∞. Furthermore, both $\mathcal{E}_{|\sigma^*(F_0)}$ and $\mathcal{F}_{|\sigma^*(F_0)}$ are pullbacks under f of vector bundles on $\widetilde{Y}$.

Proof The proof is identical to that of [19, Proposition 3.11]. □

We now describe the pullback $\sigma^*(F_1) \subseteq \widetilde{\mathfrak{G}}^5_{16}$ and we define the determinantal variety

$$(9) \qquad Z := \{(y, L) \in X \times W^5_{16}(X) : h^0(X, L(-2y)) \geq 5\}.$$

Because X is general, arguing precisely as in [19, Proposition 3.10], we find that Z is pure of dimension 2. Next we observe that in order to estimate the intersection of $[\widetilde{\mathfrak{D}}_{13}]^{\text{virt}}$ with the surface $\sigma^*(F_1)$, it suffices to restrict ourselves to Z:

Proposition 2.5 *The variety Z is an irreducible component of $\sigma^*(F_1)$, and*

$$c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)_{|\sigma^*(F_1)} = c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)_{|Z}.$$

Proof Let $(\ell_X, \ell_E) \in \sigma^{-1}([X \cup_y E])$ be a limit linear series. Observe that $\rho(13, 5, 16) = 1$, which is greater than or equal to the sum of the adjusted Brill–Noether numbers $\rho(\ell_X, y) + \rho(\ell_E, y)$; see Definition 2.1. Since $\rho(\ell_E, y) \geq 0$, it follows that $\rho(\ell_X, y) \in \{0, 1\}$. If $\rho(\ell_E, y) = 0$, then $\ell_E = 10y + |\mathcal{O}_E(6y)|$ and the aspect $\ell_X \in G^5_{16}(X)$ is a complete linear series with a cusp at the point $y \in X$. Therefore $(y, \ell_X) \in Z$, and in particular $Z \times \{\ell_E\} \cong Z$ is a union of irreducible components of $\sigma^*(F_1)$.

The remaining components of $\sigma^*(F_1)$ are indexed by Schubert indices

$$\alpha := (0 \le \alpha_0 \le \cdots \le \alpha_5 \le 11 = 16 - 5)$$

such that $\alpha \ge (0, 1, 1, 1, 1, 1)$ holds lexicographically and $\alpha_0 + \cdots + \alpha_5 \in \{6, 7\}$ when $\rho(\ell_X, y) \ge -1$ for any point $y \in X$; see also [18, Theorem 0.1]. For a Schubert index α satisfying these conditions, we let $\alpha^c := (11 - \alpha_5, \ldots, 11 - \alpha_0)$ be the complementary Schubert index, and define

$$Z_\alpha := \{(y, \ell_X) \in X \times G_{16}^5(X) : \alpha^{\ell_X}(y) \ge \alpha\} \quad \text{and} \quad W_\alpha := \{\ell_E \in G_{16}^5(E) : \alpha^{\ell_E}(y) \ge \alpha^c\}.$$

Then the following relation holds for certain natural coefficients m_α:

$$\sigma^*(F_1) = Z + \sum_{\alpha \ge (0,1,1,1,1,1)} m_\alpha (Z_\alpha \times W_\alpha).$$

We now finish the proof by invoking the pointed Brill–Noether theorem [16, Theorem 1.1], which gives $\dim Z_\alpha = 1 + \rho(12, 5, 16) - (\alpha_0 + \cdots + \alpha_5) \le 1$. In the definition of the test curve F_1, the point of attachment $y \in E$ is fixed, therefore the restrictions of both $\mathcal{E}$ and $\mathcal{F}$ are pulled-back from Z_α and one obtains $c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)_{|Z_\alpha \times W_\alpha} = 0$ for dimension reasons. □

2.4 Top Chern numbers on Jacobians

We use various facts about intersection theory on Jacobians, for which we refer to [5, Chapters VII–VIII]. We start with a general curve X of genus g, fix a Poincaré line bundle $\mathcal{P}$ on $X \times \mathrm{Pic}^d(X)$ and denote by

$$\pi_1 \colon X \times \mathrm{Pic}^d(X) \to X \text{ and } \pi_2 \colon X \times \mathrm{Pic}^d(X) \to \mathrm{Pic}^d(X)$$

the two projections. Let $\eta = \pi_1^*([x_0]) \in H^2(X \times \mathrm{Pic}^d(X), \mathbb{Z})$, where $x_0 \in X$ is a fixed point. We choose a symplectic basis $\delta_1, \ldots, \delta_{2g} \in H^1(X, \mathbb{Z}) \cong H^1(\mathrm{Pic}^d(X), \mathbb{Z})$, and then consider the class

$$\gamma := -\sum_{\alpha=1}^{g} \left(\pi_1^*(\delta_\alpha) \pi_2^*(\delta_{g+\alpha}) - \pi_1^*(\delta_{g+\alpha}) \pi_2^*(\delta_\alpha) \right) \in H^2(X \times \mathrm{Pic}^d(X), \mathbb{Z}).$$

One has $c_1(\mathcal{P}) = d \cdot \eta + \gamma$, and the relations $\gamma^3 = 0$, $\gamma\eta = 0$, $\eta^2 = 0$ and $\gamma^2 = -2\eta\pi_2^*(\theta)$, for which we refer to [5, page 335]. Assuming $W_d^{r+1}(X) = \varnothing$ (which is what happens in the case of $g = 12$, $r = 5$ and $d = 16$ relevant to us), the smooth variety $W_d^r(X)$ admits a rank-$r+1$ vector bundle

$$\mathcal{M} := (\pi_2)_* (\mathcal{P}_{|X \times W_d^r(X)})$$

with fibers $\mathcal{M}_L \cong H^0(X, L)$, for $L \in W_d^r(X)$. The Chern numbers of $\mathcal{M}$ are computed via the Harris–Tu formula [28]. We write formally

$$\sum_{i=0}^{r} c_i(\mathcal{M}^\vee) = (1 + x_1) \cdots (1 + x_{r+1}).$$

For a class $\zeta \in H^*(\mathrm{Pic}^d(X), \mathbb{Z})$, the Chern number $c_{j_1}(\mathcal{M}) \cdots c_{j_s}(\mathcal{M}) \cdot \zeta \in H^{\mathrm{top}}(W_d^r(X), \mathbb{Z})$ can be computed by repeatedly using the following formal identities:[3]

$$(10) \qquad x_1^{i_1} \cdots x_{r+1}^{i_{r+1}} \cdot \theta^{\rho(g,r,d)-i_1-\cdots-i_{r+1}} = g! \frac{\prod_{j>k}(i_k - i_j + j - k)}{\prod_{k=1}^{r+1}(g-d+2r+i_k-k)!}.$$

We now specialize to the case when X is a general curve of genus 12, thus $W_{16}^5(X)$ is a smooth 6–fold. By Grauert's Theorem, $\mathcal{N} := (R^1\pi_2)_*(\mathcal{P}_{|X \times W_{16}^5(X)})$ is locally free of rank one. Set $y_1 := c_1(\mathcal{N})$. We now explain how y_1 determines the Chern numbers of $\mathcal{M}$.

Proposition 2.6 *For a general curve X of genus 12 set $c_i := c_i(\mathcal{M}^\vee)$ for $i = 1, \ldots, 6$, and $y_1 := c_1(\mathcal{N})$. Then the following relations hold in $H^*(W_{16}^5(X), \mathbb{Z})$:*

$$c_i = \frac{\theta^i}{i!} - \frac{\theta^{i-1}}{(i-1)!} y_1 \quad \text{for } i = 1, \ldots, 6.$$

Proof For an effective divisor D of sufficiently large degree on X, there is an exact sequence

$$0 \to \mathcal{M} \to (\pi_2)_*(\mathcal{P} \otimes \mathcal{O}(\pi^*D)) \to (\pi_2)_*(\mathcal{P} \otimes \mathcal{O}(\pi_1^*D)_{|\pi_1^*D}) \to R^1\pi_{2*}(\mathcal{P}_{|X \times W_{16}^5(X)}) \to 0.$$

Recall that $\mathcal{N}$ is the vector bundle on the right in the exact sequence above. By [5, Chapter VII], we have $c_{\mathrm{tot}}\big((\pi_2)_*(\mathcal{P} \otimes \mathcal{O}(\pi_1^*D))\big) = e^{-\theta}$, and the total Chern class of $(\pi_2)_*(\mathcal{P} \otimes \mathcal{O}(\pi_1^*D)_{|\pi_1^*D})$ is trivial. We therefore, as claimed, obtain the formula

$$(1 + y_1) \cdot e^{-\theta} = 1 - c_1 + c_2 - \cdots + c_6. \qquad \square$$

Using Proposition 2.6, any Chern number on $W_{16}^5(X)$ can be expressed in terms of monomials in y_1 and θ. The following identity on $H^{12}(W_{16}^5(X), \mathbb{Z})$ follows from (10) using the canonical isomorphism $H^1(X, L) \cong H^0(X, \omega_X \otimes L^\vee)^\vee$:

$$(11) \qquad (\theta^i \cdot y_1^{6-i})_{W_{16}^5(X)} = \frac{\theta^{12}}{(12-i)!} = i! \binom{12}{i}.$$

With this preparation in place, we now compute the classes of the loci Y and Z.

Proposition 2.7 *Let $[X, q]$ be a general pointed curve of genus 12, let $\mathcal{M}$ denote the tautological rank-six vector bundle over $W_{16}^5(X)$, and set $c_i = c_i(\mathcal{M}^\vee) \in H^{2i}(W_{16}^5(X), \mathbb{Z})$ as before. Then:*

(i) $[Z] = \pi_2^*(c_5) - 6\eta\theta\pi_2^*(c_3) + (54\eta + 2\gamma)\pi_2^*(c_4) \in H^{10}(X \times W_{16}^5(X), \mathbb{Z})$.

(ii) $[Y] = \pi_2^*(c_5) - 2\eta\theta\pi_2^*(c_3) + (15\eta + \gamma)\pi_2^*(c_4) \in H^{10}(X \times W_{16}^5(X), \mathbb{Z})$.

Proof The locus Z has been defined by (9) as the degeneracy locus of a vector bundle morphism over the 7–dimensional smooth variety $X \times W_{16}^5(X)$ (observe again that $W_{16}^6(X) = \varnothing$). For each $(y, L) \in X \times W_{16}^5(X)$, there is a natural map

$$H^0(X, L \otimes \mathcal{O}_{2y})^\vee \to H^0(X, L)^\vee.$$

---

[3]See [19, Section 4.1] for a detailed discussion of how to read and apply the Harris–Tu formula in this context.

These maps viewed together induce a morphism $\zeta\colon J_1(\mathcal{P})^\vee \to \pi_2^*(\mathcal{M})^\vee$ of vector bundles. Then Z is the first degeneracy locus of ζ and applying the Porteous formula,

$$[Z] = c_5(\pi_2^*(\mathcal{M})^\vee - J_1(\mathcal{P})^\vee).$$

The Chern classes of the jet bundle $J_1(\mathcal{P})$ are computed using the standard exact sequence

$$0 \to \pi_1^*(\omega_X) \otimes \mathcal{P} \to J_1(\mathcal{P}) \to \mathcal{P} \to 0.$$

We compute the total Chern class of the formal inverse of the jet bundle as follows:

$$
\begin{aligned}
c_{\mathrm{tot}}(J_1(\mathcal{P})^\vee)^{-1} &= \left(\sum_{j\geq 0}(d(L)\eta + \gamma)^j\right)\cdot\left(\sum_{j\geq 0}((2g(X)-2+d(L))\eta + \gamma)^j\right), \\
&= (1 + 16\eta + \gamma + \gamma^2 + \cdots)\cdot(1 + 38\eta + \gamma + \gamma^2 + \cdots), \\
&= 1 + 54\eta + 2\gamma - 6\eta\theta.
\end{aligned}
$$

Multiplying this with the total class of $\pi_2^*(\mathcal{M})^\vee$, one finds the claimed formula for $[Z]$.

To compute the class of Y defined in (8), we consider the projections

$$\mu, \nu\colon X \times X \times \mathrm{Pic}^{16}(X) \to X \times \mathrm{Pic}^{16}(X),$$

and let $\Delta \subseteq X \times X \times \mathrm{Pic}^{16}(X)$ be the diagonal. Set $\Gamma_q := \{q\} \times \mathrm{Pic}^{16}(X)$ and consider the vector bundle $\mathcal{B} := \mu_*(\nu^*(\mathcal{P}) \otimes \mathcal{O}_{\Delta + \nu^*(\Gamma_q)})$. There is a morphism $\chi\colon \mathcal{B}^\vee \to (\pi_2)^*(\mathcal{M})^\vee$ of vector bundles over $X \times W_{16}^5(X)$ obtained as the dual of the evaluation map, and the surface Y is realized as its degeneracy locus. Since we also have that

$$c_{\mathrm{tot}}(\mathcal{B}^\vee)^{-1} = \big(1 + (d(L)\eta + \gamma) + (d(L)\eta + \gamma)^2 + \cdots\big)\cdot(1 - \eta) = 1 + 15\eta + \gamma - 2\eta\theta,$$

we find the stated expression for $[Y]$ and finish the proof. $\qquad\square$

We introduce two further vector bundles which appear in many of our calculations. Their Chern classes are computed via Grothendieck–Riemann–Roch.

Proposition 2.8 *Let $[X, q]$ be a general pointed curve of genus* 12 *and consider the vector bundles* $\mathcal{A}_2$ *and* $\mathcal{B}_2$ *on* $X \times \mathrm{Pic}^{16}(X)$ *having fibers*

$$\mathcal{A}_{2,(y,L)} = H^0(X, L^{\otimes 2}(-2y)) \quad \text{and} \quad \mathcal{B}_{2,(y,L)} = H^0(X, L^{\otimes 2}(-y-q)),$$

respectively. One then has the following formulas for their Chern classes:

$$
\begin{aligned}
c_1(\mathcal{A}_2) &= -4\theta - 4\gamma - 86\eta, & c_1(\mathcal{B}_2) &= -4\theta - 2\gamma - 31\eta, \\
c_2(\mathcal{A}_2) &= 8\theta^2 + 320\eta\theta + 16\gamma\theta, & c_2(\mathcal{B}_2) &= 8\theta^2 + 116\eta\theta + 8\theta\gamma.
\end{aligned}
$$

Proof We apply Grothendieck–Riemann–Roch to the projection map

$$\nu\colon X \times X \times \mathrm{Pic}^{16}(X) \to X \times \mathrm{Pic}^{16}(X).$$

Via Grauert's theorem, $\mathcal{A}_2$ can be realized as a pushforward under the map ν, precisely

$$\mathcal{A}_2 = \nu_!\big(\mu^*(\mathcal{P}^{\otimes 2} \otimes \mathcal{O}_{X \times X \times \mathrm{Pic}^{16}(X)}(-2\Delta))\big) = \nu_*\big(\mu^*(\mathcal{P}^{\otimes 2} \otimes \mathcal{O}_{X \times X \times \mathrm{Pic}^{16}(X)}(-2\Delta))\big).$$

Applying Grothendieck–Riemann–Roch to ν, we find $\mathrm{ch}_2(\mathcal{A}_2) = 8\eta\theta$, and $\nu_*(c_1(\mathcal{P})^2) = -2\theta$. One then obtains $c_1(\mathcal{A}_2) = -4\theta - 4\gamma - (4d(L) + 2g(C) - 2)\eta$, which yields the formula for $c_2(\mathcal{A}_2)$. To determine the Chern classes of $\mathcal{B}_2$, we observe $c_1(\mathcal{B}_2) = -4\theta - 2\gamma - (2d-1)\eta$ and $\mathrm{ch}_2(\mathcal{B}_2) = 4\eta\theta$. $\qquad\square$

3 The class of the virtual divisor $\widetilde{\mathfrak{D}}_{13}$

In this section we determine the virtual class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}} := \sigma_*\big(c_2(\mathrm{Sym}^2(\mathcal{E}))^\vee - \mathcal{F}^\vee\big)$ on $\widetilde{\mathcal{M}}_{13}$. We begin by recording the following formulas for a vector bundle $\mathcal{V}$ of rank $r+1$ on a stack $\mathcal{X}$:

$$c_1(\mathrm{Sym}^2(\mathcal{V})) = (r+2)c_1(\mathcal{V}) \quad \text{and} \quad c_2(\mathrm{Sym}^2(\mathcal{V})) = \tfrac{1}{2}r(r+3)c_1^2(\mathcal{V}) + (r+3)c_2(\mathcal{V}).$$

We apply these formulas for the first degeneracy locus of $\phi^\vee \colon \mathcal{F}^\vee \to \mathrm{Sym}^2(\mathcal{E})^\vee$. By Definition 2.2, its class $[\mathfrak{U}]^{\mathrm{virt}}$ is given by

$$\text{(12)} \qquad c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) = c_2(\mathrm{Sym}^2(\mathcal{E})^\vee) - c_1(\mathrm{Sym}^2(\mathcal{E})^\vee) \cdot c_1(\mathcal{F}^\vee) + c_1^2(\mathcal{F}^\vee) - c_2(\mathcal{F}^\vee)$$

$$= 20c_1^2(\mathcal{E}) + 8c_2(\mathcal{E}) - 7c_1(\mathcal{E}) \cdot c_1(\mathcal{F}) + c_1^2(\mathcal{F}) - c_2(\mathcal{F}).$$

In what follows we expand the virtual class in $CH^1(\widetilde{\mathcal{M}}_{13})$ as

$$\text{(13)} \qquad\qquad\qquad [\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}} = a\lambda - b_0\delta_0 - b_1\delta_1.$$

We compute the coefficients a, b_0 and b_1, by intersecting both sides of this expression with the test curves F_0, F_1 and F_{ell}. We start with the coefficient b_1.

Theorem 3.1 *Let X be a general curve of genus* 12. *The coefficient b_1 in* (13) *is*

$$b_1 = \frac{1}{2g(X) - 2}\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) = 11787.$$

Proof We intersect the degeneracy locus of the map $\phi \colon \mathrm{Sym}^2(\mathcal{E}) \to \mathcal{F}$ with $\sigma^*(F_1)$. By Proposition 2.5, it suffices to estimate the contribution coming from Z. We write

$$\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) = c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)_{|Z}.$$

In Proposition 2.7, we constructed a morphism $\zeta \colon J_1(\mathcal{P})^\vee \to \pi_2^*(\mathcal{M})^\vee$ of vector bundles on Z, whose fibers are the maps $H^0(\mathcal{O}_{2y})^\vee \to H^0(X, L)^\vee$. The kernel sheaf $\mathrm{Ker}(\zeta)$ is locally free of rank one. If U is the line bundle on Z with fiber

$$U(y, L) = \frac{H^0(X, L)}{H^0(X, L(-2y))} \hookrightarrow H^0(X, L \otimes \mathcal{O}_{2y})$$

over a point $(y, L) \in Z$, then over Z one has the exact sequence

$$0 \to U \to J_1(\mathcal{P}) \to (\mathrm{Ker}(\zeta))^\vee \to 0.$$

In particular, by Proposition 2.7, we find that

$$(14) \qquad\qquad c_1(U) = 2\gamma + 54\eta + c_1(\mathrm{Ker}(\zeta)).$$

The product of the Chern class of $\mathrm{Ker}(\zeta)$ with any class $\xi \in H^2(X \times W_{16}^5(X), \mathbb{Z})$ is given by the Harris–Tu formula [28]

$$(15)\ \ c_1(\mathrm{Ker}(\zeta))\cdot\xi_{|Z} = -c_6(\pi_2^*(\mathcal{M})^\vee - J_1(\mathcal{P})^\vee)\cdot\xi_{|Z} = -\big(\pi_2^*(c_6) - 6\eta\theta\pi_2^*(c_4) + (54\eta + 2\gamma)\pi_2^*(c_5)\big)\cdot\xi_{|Z}.$$

Similarly, one has the formula [28] for the self-intersection on the surface Z:

$$(16) \qquad c_1^2(\mathrm{Ker}(\zeta)) = \big(\pi_2^*(c_7) - 6\eta\theta\pi_2^*(c_5) + (54\eta + 2\gamma)\pi_2^*(c_6)\big) \in H^{14}(X \times W_{16}^5(X), \mathbb{Z}) \cong \mathbb{Z}.$$

We also observe that $c_7 = 0$, since the bundle $\mathcal{M}$ has rank 6.

Let $\mathcal{A}_3$ denote the vector bundle on Z having fibers

$$\mathcal{A}_{3,(y,L)} = H^0(X, L^{\otimes 2})$$

constructed as a pushforward of a line bundle on $X \times X \times \mathrm{Pic}^{16}(X)$. Then the line bundle $U^{\otimes 2}$ can be embedded in $\mathcal{A}_3/\mathcal{A}_2$. We consider the quotient

$$\mathcal{G} := \frac{\mathcal{A}_3/\mathcal{A}_2}{U^{\otimes 2}}.$$

The morphism $U^{\otimes 2} \to \mathcal{A}_3/\mathcal{A}_2$ vanishes along the locus of pairs (y, L), where L has a basepoint. It follows that the sheaf $\mathcal{G}$ has torsion along the locus $\Gamma \subseteq Z$ consisting of pairs $(q, A(q))$, where $A \in W_{16}^5(X)$. Furthermore, $\mathcal{F}_{|Z}$, as a subsheaf of $\mathcal{A}_3$, can be identified with the kernel of the map $\mathcal{A}_3 \to \mathcal{G}$. Summarizing, there is an exact sequence of vector bundles on Z,

$$(17) \qquad\qquad 0 \to \mathcal{A}_{2|Z} \to \mathcal{F}_{|Z} \to U^{\otimes 2} \to 0.$$

Over a general point $(y, L) \in Z$, this sequence reflects the decomposition

$$\mathcal{F}(y, L) = H^0(X, L^{\otimes 2}(-2y)) \oplus K \cdot u^2,$$

where $u \in H^0(X, L)$ is a section such that $\mathrm{ord}_y(u) = 1$.

Via the exact sequence (17), one computes the Chern classes of $\mathcal{F}_{|Z}$:

$$c_1(\mathcal{F}_{|Z}) = c_1(\mathcal{A}_{2|Z}) + 2c_1(U) \quad \text{and} \quad c_2(\mathcal{F}_{|Z}) = c_2(\mathcal{A}_{2|Z}) + 2c_1(\mathcal{A}_{2|Z})c_1(U).$$

Recalling that $\mathcal{E}_{|Z} = \pi_2^*(\mathcal{M})_{|Z}$ and using (12), we find that $\sigma^*(F_1) \cdot c_2((\mathrm{Sym}^2 \mathcal{E})^\vee - \mathcal{F}^\vee)$ is equal to

$$20c_1^2(\pi_2^*\mathcal{M}_{|Z}^\vee) + 8c_2(\pi_2^*\mathcal{M}_{|Z}^\vee) + 7c_1(\pi_1^*\mathcal{M}_{|Z}^\vee)\cdot c_1(\mathcal{A}_{2|Z}) + 4c_1^2(U)$$
$$- c_2(\mathcal{A}_{2|Z}) + 14c_1(\pi_2^*\mathcal{M}_{|Z}^\vee)\cdot c_1(U) + c_1^2(\mathcal{A}_{2|Z}) + 2c_1^2(\mathcal{A}_{2|Z})\cdot c_1(U).$$

Here, $c_i(\pi_2^*\mathcal{M}_{|Z}^\vee) = \pi_2^*(c_i) \in H^{2i}(Z, \mathbb{Z})$. The Chern classes of $\mathcal{A}_{2|Z}$ were computed in Proposition 2.8. Formula (14) expresses $c_1(U)$ in terms of $c_1(\mathrm{Ker}(\zeta))$ and the classes η and γ. When expanding $\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)$, one distinguishes between terms that *do* and those that *do not* contain

the first Chern class of $\mathrm{Ker}(\zeta)$. The coefficient of $c_1(\mathrm{Ker}(\zeta))$, as well as the contribution coming from $c_1^2(\mathrm{Ker}(\zeta))$ in the expression of $\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)$ is evaluated using the formulas (15) and (16) respectively. To carry this out, we first consider the part of this product that *does not* contain $c_1(\mathrm{Ker}(\zeta))$, and we obtain

$$8\pi_2^*(c_2) + 20\pi_2^*(c_1^2) + c_1^2(\mathcal{A}_{2|Z}) + 7\pi_2^*(c_1) \cdot c_1(\mathcal{A}_{2|Z}) - c_2(\mathcal{A}_{2|Z}) + 4(2\gamma + 54\eta)^2$$
$$+ 2(2\gamma + 54\eta) \cdot c_1(\mathcal{A}_{2|Z}) + 14(2\gamma + 54\eta) \cdot \pi_2^*(c_1)$$
$$= 20\pi_2^*(c_1^2) + 154\pi_2^*(c_1) \cdot \eta - 28\pi_2^*(c_1) \cdot \theta - 96\eta\theta + 8\theta^2 + 8\pi_2^*(c_2)$$

in $H^4(X \times W_{16}^5(X), \mathbb{Z})$. This polynomial gets multiplied by the class $[Z]$, which is expressed in Proposition 2.7 as a degree 5 polynomial in θ, η and $\pi_2^*(c_i)$. We obtain a homogeneous polynomial of degree 7 viewed as an element of $H^{14}(X \times W_{16}^5(X), \mathbb{Z})$.

Next we turn our attention to the contribution $\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)$ coming from terms that do contain $c_1(\mathrm{Ker}(\zeta))$. This is given by the formula

$$4c_1^2(\mathrm{Ker}(\zeta)) + c_1(\mathrm{Ker}(\zeta)) \cdot \big(8(2\gamma + 54\eta) + 2c_1(\mathcal{A}_{2|Z}) + 14\pi_2^*(c_1)\big).$$

Using (15) and (16), one ends up with the following homogeneous polynomial of degree 7 in η, θ and $\pi_2^*(c_i)$ for $i = 1, \ldots, 6$:

$$84\pi_2^*(c_1 c_4)\theta\eta - 48\pi_2^*(c_4)\theta^2\eta - 756\pi_2^*(c_1 c_5)\eta + 440\pi_2^*(c_5)\theta\eta - 44\pi_2^*(c_6)\eta.$$

Adding together the parts that do and those that do not contain $c_1(\mathrm{Ker}(\zeta))$, and using the fact that the only monomials that need to be retained are those containing η, after manipulations carried out using *Maple*, one finds

$$\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)$$
$$= \eta\pi_2^*\big(-602c_1 c_5 + 432c_2 c_4 - 120c_1^2 c_3\theta + 168c_1 c_3\theta^2 - 48c_3\theta^3 + 1080c_1^2 c_4 - 1428c_1 c_4\theta$$
$$- 48c_2 c_3\theta + 384c_4\theta^2 + 344c_5\theta - 44c_6\big).$$

We suppress η and the remaining polynomial lives inside $H^{12}(W_{16}^5(X), \mathbb{Z}) \cong \mathbb{Z}$. Using Proposition 2.6 this expression is equal to

$$\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) = \tfrac{193}{45}\theta^6 - \tfrac{1271}{30}\theta^5 y_1 + \tfrac{1607}{12}\theta^4 y_1^2 - 120\theta_3 y_1^3 = 259314,$$

where for the last step we used the formulas (11). We conclude

$$b_1 = \tfrac{1}{22}\sigma^*(F_1) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) = 11787,$$

as required. $\qquad\square$

Theorem 3.2 *Let $[X, q]$ be a general pointed curve of genus 12 and let $F_0 \subseteq \widetilde{\Delta}_0 \subseteq \widetilde{\mathcal{M}}_{13}$ be the associated test curve. Then the coefficient of δ_0 in the expression (13) of $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$ is equal to*

$$b_0 = \tfrac{1}{24}\big(\sigma^*(F_0) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) + b_1\big) = 2247.$$

Proof Using Proposition 2.4, we observe that

$$c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)_{|\sigma^*(F_0)} = c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)_{|Y}.$$

We shall evaluate the Chern classes of $\mathcal{F}_{|Y}$ via the line bundle V on Y with fiber

$$V(y, L) = \frac{H^0(X, L)}{H^0(X, L(-y-q))} \hookrightarrow H^0(X, L \otimes \mathcal{O}_{y+q})$$

over a point $(y, L) \in Y$. We write the exact sequence

$$0 \to V \to \mathcal{B} \to (\mathrm{Ker}(\chi))^\vee \to 0$$

over Y, where the morphism $\chi: \mathcal{B}^\vee \to \pi_2^*(\mathcal{M})^\vee$ was defined in the final part of the proof of Proposition 2.7. In particular, we have

$$c_1(V) = 15\eta + \gamma + c_1(\mathrm{Ker}(\chi)).$$

The effect of multiplying $c_1(\mathrm{Ker}(\chi))$ against a class $\xi \in H^2(X \times W_{16}^5(X), \mathbb{Z})$ is described by applying once more the Harris–Tu formula [28]:

$$(18) \qquad c_1(\mathrm{Ker}(\chi)) \cdot \xi_{|Y} = \left(-\pi_2^*(c_6) - 2\eta\theta\pi_2^*(c_4) + (15\eta + \gamma)\pi_2^*(c_5)\right) \cdot \xi_{|Y},$$

where we recall that $\pi_2: X \times W_{16}^5(X) \to W_{16}^5(X)$ and $c_i \in H^{2i}(W_{16}^5(X), \mathbb{Z})$. Similarly, for the self-intersection on Y the following formula holds:

$$(19) \qquad c_1^2(\mathrm{Ker}(\chi)) = -2\eta\theta\pi_2^*(c_5) + (15\eta + \gamma)\pi_2^*(c_6) \in H^{14}(X \times W_{16}^5(X), \mathbb{Z}).$$

We have also introduced in Proposition 2.8 the vector bundle $\mathcal{B}_2$ on $X \times \mathrm{Pic}^{16}(X)$ with fibers $\mathcal{B}_{2,(y,L)} = H^0(X, L^{\otimes 2}(-y-q))$ over a point (y, L). A local calculation along the lines of the one in the proof of Theorem 3.1 shows that one also has an exact sequence on Y, which can then be used to determine the Chern numbers of $\mathcal{F}_{|Y}$:

$$0 \to \mathcal{B}_{2|Y} \to \mathcal{F}_{|Y} \to V^{\otimes 2} \to 0.$$

This exact sequence reflects the fact for a general point $(y, L) \in Y$ one has a decomposition $\mathcal{F}(y, L) = H^0(X, L^{\otimes 2}(-y-q)) \oplus K \cdot u^2$, where $u \in H^0(X, L)$ is a section that does not vanish at y and q. We thus obtain the formulas

$$c_1(\mathcal{F}_{|Y}) = c_1(\mathcal{B}_{2|Z}) + 2c_1(V) \quad \text{and} \quad c_2(\mathcal{F}_{|Y}) = c_2(\mathcal{B}_{2|Y}) + 2c_1(\mathcal{B}_{2|Y})c_1(V).$$

To estimate $c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee)_{|Y}$ we use (12) and write

$$\begin{aligned}
\sigma^*(F_0) \cdot c_2((\mathrm{Sym}^2 \mathcal{E})^\vee - \mathcal{F}^\vee) \\
= 20c_1^2(\pi_1^* \mathcal{M}_{|Y}^\vee) + 8c_2(\pi_2^* \mathcal{M}_{|Y}^\vee) + 7c_1(\pi_1^* \mathcal{M}_{|Y}^\vee) \cdot c_1(\mathcal{B}_{2|Y}) + 4c_1^2(V) - c_2(\mathcal{B}_{2|Y}) \\
+ 14c_1(\pi_2^* \mathcal{M}_{|Y}^\vee) \cdot c_1(V) + c_1^2(\mathcal{B}_{2|Y}) + 2c_1(\mathcal{B}_{2|Y}) \cdot c_1(V).
\end{aligned}$$

We expand this expression, collect the terms that do not contain $c_1(\mathrm{Ker}(\chi))$, and obtain

$$20\pi_2^*(c_1^2) - 7\eta\pi_2^*(c_1) - 28\theta \cdot \pi_2^*(c_1) + 4\theta\eta + 8\theta^2 + 8\pi_2^*(c_2).$$

This quadratic polynomial gets multiplied with the class $[Y]$ computed in Proposition 2.7. Next, we collect the terms in $\sigma^*(F_0) \cdot c_2(\mathrm{Sym}^2 \mathcal{E}^\vee - \mathcal{F}^\vee)$ that do contain $c_1(\mathrm{Ker}(\chi))$:

$$4c_1^2(\mathrm{Ker}(\chi)) + c_1(\mathrm{Ker}(\chi))\big(8(15\eta + \gamma) + 14\pi_2^*(c_1) + 2c_1(\mathcal{B}_{2|Y})\big).$$

This part of the contribution is evaluated using formulas (18) and (19).

Putting everything together, we obtain a polynomial in $H^{14}(X \times W_{16}^5(X), \mathbb{Z}) \cong \mathbb{Z}$, as in the proof of Theorem 3.1:

$$
\begin{aligned}
\sigma^*&(F_0) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) \\
&= \eta\pi_2^*\big(-40c_1^2 c_3\theta + 56c_1 c_3\theta^2 - 16c_3\theta^3 + 300c_1^2 c_4 - 392c_1 c_4\theta - 16c_2 c_3\theta + 104c_4\theta^2 - 217c_1 c_5 \\
&\qquad\qquad + 120c_2 c_4 + 124c_5\theta + 2c_6\big).
\end{aligned}
$$

Applying Proposition 2.6 and then (11), after eliminating η we obtain

$$\sigma^*(F_0) \cdot c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) = \tfrac{161}{180}\theta^6 - \tfrac{28}{3}\theta^5 y_1 + \tfrac{755}{24}\theta^4 y_1^2 - 30\theta^3 y_1^3 = 42141. \qquad \square$$

We can now complete the calculation of $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$.

Proof of Theorem 1.4 We consider the curve $F_{\mathrm{ell}} \subseteq \widetilde{\mathcal{M}}_g$ defined in (7) obtained by attaching at the fixed point of a general curve X of genus 12 a pencil of plane cubics at one of the basepoints of the pencil. Then one has the relation

$$a - 12b_0 + b_1 = F_{\mathrm{ell}} \cdot \sigma_* c_2(\mathrm{Sym}^2(\mathcal{E})^\vee - \mathcal{F}^\vee) = 0.$$

Using Theorems 3.1 and 3.2, we thus find $a = 15177$ for the λ–coefficient in the expansion (13). This completes the calculation of the virtual class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$. $\qquad \square$

We finally explain how Theorems 1.4 and 1.5 (proved in Section 4) together imply Theorem 1.3.

Proof of Theorem 1.3 We write $[\overline{\mathfrak{D}}_{13}] = a\lambda - b_0\delta_0 - \cdots - b_6\delta_6$, where a, b_0 and b_1 are determined by Theorem 1.4. Applying [21, Theorem 1.1], we have the inequalities $b_i \geq (6i + 8)b_0 - (i + 1)a \geq b_0$ for $i = 2, \ldots, 6$, which shows that $s(\overline{\mathfrak{D}}_{13}) = a/b_0 = \tfrac{5059}{749}$. $\qquad \square$

4 The strong maximal rank conjecture in genus 13

In this section and the next, we prove that $\widetilde{\mathfrak{D}}_{13}$ is not all of $\widetilde{\mathcal{M}}_{13}$ and that its codimension-one part represents the virtual class $[\widetilde{\mathfrak{D}}_{13}]^{\mathrm{virt}}$.

To show that $\widetilde{\mathfrak{D}}_{13}$ is not all of $\widetilde{\mathcal{M}}_{13}$, it suffices to prove the existence of one Brill–Noether general smooth curve X of genus 13 such that, for every $L \in W_{16}^5(X)$, the multiplication map

$$\phi_L : \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$$

is surjective. This is one case of the strong maximal rank conjecture [3]. The locus of such curves is Zariski open; to prove that it is nonempty over every algebraically closed field of characteristic zero, it suffices to show this over one such field. Hence, we can and do assume that our ground field K is spherically complete with respect to a surjective valuation $\nu \colon K^\times \to \mathbb{R}$, and that K has residue characteristic zero. This allows us to discuss the nonarchimedean analytifications of curves, the skeletons of those analytifications, and the tropicalizations of rational functions, viewed as sections of L and $L^{\otimes 2}$. In this framework, we apply the method of tropical independence to give a lower bound for the rank of the multiplication map ϕ_L for every $L \in W^5_{16}(X)$. The motivation and technical foundations for this approach are detailed in Sections 1.4–1.5, Sections 2.4–2.5 and Section 6 of [19], to which we refer the reader for details and further references.

After proving this case of the strong maximal rank conjecture, we will furthermore show that no component of the degeneracy locus $\mathfrak{U}$ in the parameter space $\widetilde{\mathfrak{G}}^5_{16}$ over $\widetilde{\mathcal{M}}_{13}$ maps with generically positive-dimensional fibers onto a divisor in $\widetilde{\mathcal{M}}_{13}$. As in [19], this additional step is necessary to show that the pushforward of the virtual class is effective, and our proof involves analogous arguments on lower-genus curves for linear series with ramification. In particular, we will consider linear series with ramification on curves of genus 11 and 12 in Section 5, and so we set up our arguments here to work in this greater generality.

Let X be a smooth projective curve of genus $11 \le g \le 13$ over K whose Berkovich analytification X^{an} has a skeleton Γ which is a chain of g loops connected by bridges, as shown. In order to simplify notation later, the vertices of Γ are labeled $w_{13-g}, \ldots, w_{13}$, and $v_{14-g}, \ldots, v_{14}$, as shown in Figure 1.

For $14 - g \le k \le 13$ we write γ_k for the loop formed by the two edges of length ℓ_k and m_k between v_k and w_k. Similarly, for $14 - g \le k \le 14$ we write β_k for the bridge between w_{k-1} and v_k which has length n_k. Except where stated otherwise, we assume that these edge lengths satisfy

$$(20) \qquad\qquad \ell_{k+1} \ll m_k \ll \ell_k \ll n_{k+1} \ll n_k \quad \text{for all } k.$$

These conditions on the edge lengths are precisely as in [19, Section 7.1]. Any curve X whose analytification has such a skeleton is Brill–Noether general [12].

Given a line bundle L on X we choose an identification $L = \mathcal{O}_X(D_X)$ so that any linear series $V \subseteq H^0(X, L)$ is identified with a finite-dimensional vector space of rational functions $V \subseteq K(X)$. The tropicalization of any nonzero rational function f on X is a piecewise linear function with integer slopes on Γ, and we write trop V for the set of all tropicalizations of nonzero functions in V.
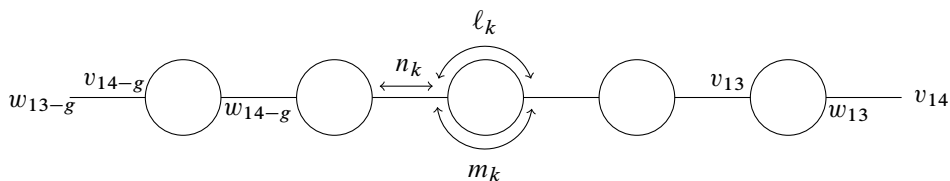

Figure 1: The chain of loops Γ.

Any sum of two functions in trop V is the tropicalization of a function in the image of the multiplication map $\phi_V \colon \mathrm{Sym}^2 V \to H^0(X, L^{\otimes 2})$. We say that a set of functions $\{\psi_0, \ldots, \psi_n\}$ on Γ is *tropically independent* if there are real numbers $b_0, \ldots, b_n$ such that

$$\min\{\psi_0 + b_0, \ldots, \psi_n + b_n\} \neq \min\{\psi_0 + b_0, \ldots, \widehat{\psi_j + b_j}, \ldots, \psi_n + b_n\} \quad \text{for } 0 \leq j \leq n.$$

In other words, $\{\psi_0, \ldots, \psi_n\}$ is tropically independent if there are real numbers $b_0, \ldots, b_n$ such that each $\psi_j + b_j$ achieves the minimum uniquely in $\min_i\{\psi_i + b_i\}$ at some point $v \in \Gamma$. The function $\theta = \min_i\{\psi_i + b_i\}$ is then called an *independence*, since it verifies that $\{\psi_0, \ldots, \psi_n\}$ is independent.

We recall that tropical independence is a sufficient condition for linear independence; if $f_0, \ldots, f_n$ are nonzero rational functions on X such that $\{\mathrm{trop}(f_0), \ldots, \mathrm{trop}(f_n)\}$ is tropically independent on Γ, then $\{f_0, \ldots, f_n\}$ is linearly independent in $K(X)$. Therefore, the relevant case of the strong maximal rank conjecture, and hence the fact that $\widetilde{\mathfrak{D}}_{13}$ is a divisor, follows immediately from the following.

Theorem 4.1 *Let X be a curve of genus 13 with skeleton Γ. Let V be a linear series of degree 16 and dimension 5 on X, and let $\Sigma = \mathrm{trop}\, V$. Then there is an independence θ among 20 pairwise sums of functions in Σ.*

We will use the following generalization of Theorem 4.1 in our proof that $\widetilde{\mathfrak{D}}_{13}$ represents the virtual class; the generalization involves analogous statements for linear series satisfying certain ramification conditions in genus 11 and 12. The situation is closely parallel to that in [19, Section 9.4]. Recall that $a_0^V(p) < \cdots < a_r^V(p)$ denotes the vanishing sequence of a linear series V of rank r at a point p.

Theorem 4.2 *Let X be a curve of genus $g \in \{11, 12, 13\}$ whose skeleton is Γ, and let $p \in X$ be a point specializing to w_{13-g}. Let V be a linear series of degree 16 and dimension 5 on X, and let $\Sigma = \mathrm{trop}\, V$. Assume that*

   (i)   *if $g = 12$, then $a_1^V(p) \geq 2$, and*

   (ii)   *if $g = 11$, then either $a_1^V(p) \geq 3$, or $a_0^V(p) \geq 1$ and $a_2^V(p) \geq 4$.*

Then there is an independence θ among 20 pairwise sums of functions in Σ.

The remainder of this section is devoted to the proof of Theorem 4.2. Our approach to constructing the independence is similar to that of [19], with a few important differences that we highlight when they arise. Throughout, we let D_X be a divisor class on X with $V \subseteq H^0(X, \mathcal{O}(D_X))$. We write $D = \mathrm{Trop}(D_X)$, and we assume that D is a break divisor, meaning that it is the unique effective representative of its equivalence class with multiplicity $\deg D - g$ at w_0 and precisely one point of multiplicity 1 on each loop γ_k. (See for instance [2].) Let $R(D)$ denoted the complete tropical linear series of D, as in [25]. In other words, $R(D) = \{\psi \in \mathrm{PL}(\Gamma) : D + \mathrm{div}(\psi) \geq 0\}$. Note, in particular, that $\mathrm{Trop}(V)$ is a tropical submodule of $R(D)$.

Remark 4.3 The differences between the constructions of independences here and those in [19] are subtle but crucial. Even when $g = 13$, $[D]$ is vertex-avoiding, and Σ is unramified (the cases treated in Section 4.1), if we apply the algorithm of [19, Section 8.1] naively, we obtain an independence among only 19 functions in Σ. To overcome this difficulty, we divide the graph into blocks in such a way that the lingering loop is the last loop in its block and has exactly two permissible functions. This allows us to alter the algorithm slightly and assign a function to the lingering loop, raising the total number of functions in the independence to 20. See Remark 4.11.

4.1 The unramified vertex-avoiding case

We first consider the case where $g = 13$, D is vertex-avoiding, and $\Sigma = \text{trop } V$ is unramified. Unramified means that the ramification weights of trop V at w_0 and v_{14}, in the sense of [19, Definition 9.7], are zero. Vertex-avoiding means that, for $0 \leq i \leq 5$, there is a unique divisor $D_i \sim D$ such that $D_i - i w_0 - (5-i) v_{14}$ is effective. A vertex-avoiding divisor is unramified if and only if the support of $D_i - i w_0 - (5-i) v_{14}$ contains neither w_0 nor v_{14}, for all i.

For $\psi \in \Sigma$, we write $s_k(\psi)$ and $s'_k(\psi)$ for the rightward slopes along the incoming and outgoing bridges of the kth loop γ_k, at v_k and w_k, respectively. Since $\dim V = 6$, the functions in Σ have exactly 6 distinct slopes along each tangent vector in Γ.

Definition 4.4 Let $s_k[0] < \cdots < s_k[5]$ and $s'_k[0] < \cdots < s'_k[5]$ denote the 6 distinct rightward slopes that occur as $s_k(\psi)$ and $s'_k(\psi)$ for $\psi \in \Sigma$.

Since D is vertex-avoiding, there is a function $\varphi_i \in \Sigma$ such that

$$s_k(\varphi_i) = s_k[i] \quad \text{and} \quad s'_k(\varphi_i) = s'_k[i] \quad \text{for all } k,$$

and it is unique up to additive constants. Since Σ is also unramified, there is a unique *lingering loop* γ_ℓ, ie a unique loop γ_ℓ such that $s'_\ell[i] = s_\ell[i]$ for all i. Moreover, there is no function $\varphi \in \Sigma$ with the property that $s_\ell(\varphi) \leq s_\ell[i]$ and $s'_\ell(\varphi) \geq s'_\ell[i+1]$. This last condition means that γ_ℓ is not a *switching loop*, in the sense of [19, Section 9.6].

Our assumption that Σ is unramified implies that the break divisor D satisfies $\deg_{w_0} D = 3$, and the rightward slopes of the functions ψ_i at w_0 are

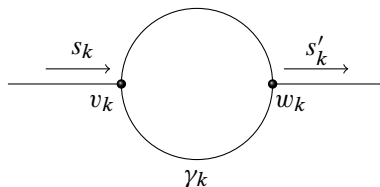$$(s'_0[0], \ldots, s'_0[5]) = (-2, -1, 0, 1, 2, 3).$$

Figure 2: The slopes s_k and s'_k.

Let us consider how the slope vector $(s'_k[0], \ldots, s'_k[5])$ changes as we go from left to right across the graph. When crossing a loop other than the lingering loop γ_ℓ, one of these slopes increases by 1, and the other 5 remain the same. So, after the first nonlingering loop, the slopes are $(-2, -1, 0, 1, 2, 4)$, and after the second nonlingering loop, the slopes are either $(-2, -1, 0, 1, 2, 5)$ or $(-2, -1, 0, 1, 3, 4)$. The data of these slopes is recorded by a standard Young tableau on a rectangle with 2 rows and 6 columns, filled with the symbols 1 through 13, excluding ℓ. If the symbol k appears in column i, then it is the $(5-i)^{\text{th}}$ slope that increases on the loop γ_k, ie $s'_k[5-i] = s_k[5-i] + 1$. Note, in particular, that each slope increases exactly twice, so $s'_{13} = (0, 1, 2, 3, 4, 5)$ and no slope is ever greater than 5.

Let $\varphi_{ij} := \varphi_i + \varphi_j$. To prove Theorem 4.2, we construct an independence θ among 20 of the 21 functions in

$$\mathcal{B} = \{\varphi_{ij} : 0 \leq i \leq j \leq 5\}.$$

In order to describe this construction, we divide the graph into three connected regions consisting of some number of loops and the bridges between them, which we call *blocks*. The construction ensures that, within each block, the slope of θ is nearly constant on each bridge, equal to 4 on bridges in the first block, 3 on bridges in the second block, and 2 on bridges in the third block. The slope decreases by 1 at the midpoint of the bridges between blocks.

The blocks are specified as follows. Recall that γ_ℓ is the lingering loop. Let

$$z_1 = \min\{6, \ell\} \quad \text{and} \quad z_2 = \max\{7, \ell\}.$$

Then γ_{z_1} and γ_{z_2} are the last loops of the first and second blocks, respectively. We construct our independence θ to satisfy

(21)
$$s_k(\theta) = \begin{cases} 4 & \text{if } k \leq z_1, \\ 3 & \text{if } z_1 < k \leq z_2, \\ 2 & \text{if } z_2 < k \leq 13. \end{cases}$$

Note that either z_1 or z_2 is equal to ℓ, so the lingering loop γ_ℓ is always the last loop in its block.

When we construct θ as a tropical linear combination of the functions in $\mathcal{B}$, we keep track of which functions achieve the minimum on which loops and bridges of Γ. The specified slopes of θ along the bridges within each block place natural constraints on which functions can achieve the minimum on a given loop, which we encode in the following definition of *permissibility*. In the vertex-avoiding case, we apply this condition only to functions $\varphi_{ij} \in \mathcal{B}$. However, we state the definition of permissibility more generally, for arbitrary functions ψ in the complete tropical linear series $R(D)$, for later use in Sections 4.2–4.3.

Definition 4.5 Let $\psi \in R(D)$. We say that ψ is *permissible* on γ_k if

$$s_k(\psi) \leq s_k(\theta) \quad \text{and} \quad s'_k(\psi) \geq s_k(\theta).$$

We say that ψ is *permissible* on a block if it is permissible on some loop in that block.

To understand this definition, suppose that θ has nearly constant slope along the bridges within each block and on each half of the bridges between blocks, and that it is written as the minimum of finitely many functions in $R(D)$, including ψ. If $s_k(\psi) \geq s_k(\theta) + 1$, then the value of ψ at v_k exceeds the value of θ at v_k by at least the length of the bridge β_k (or half this length, if β_k is the bridge between two blocks). Since this bridge is much longer than the loop γ_k, it follows that ψ cannot achieve the minimum at any point of γ_k. A similar argument shows that if $s'_k(\psi) \leq s_k(\theta) - 1$, then ψ cannot achieve the minimum at any point of γ_k.

We construct θ algorithmically, moving from left to right across the graph. At each step, we keep track of which functions in $\mathcal{B}$ are permissible on the given loop. The set of loops on which a given function ψ is permissible are indexed by the integers in an interval [19, page 44], so we pay special attention to the first and last loops in these intervals.

Suppose γ_k is the first loop on which $\varphi_{ij} \in \mathcal{B}$ is permissible and it is not the first loop in its block. Then γ_k is the unique loop on which φ_{ij} is permissible such that the first inequality in Definition 4.5 is strict. Similarly, suppose γ_k is the last loop on which φ_{ij} is permissible and it is not the last loop in its block. Then γ_k is the unique loop on which φ_{ij} is permissible such that the second inequality in Definition 4.5 is strict. This motivates the following definition.

Definition 4.6 A permissible function ψ is *new* if $s_k(\psi) \leq s_k(\theta) - 1$ and *departing* if $s'_k(\psi) \geq s_k(\theta) + 1$.

Our choice of z_1 and z_2 determines which loops have new permissible functions in $\mathcal{B}$.

Proposition 4.7 *There are no new permissible functions of the form φ_{ij} on γ_k if and only if $k = \ell$ or*

(i) $\ell > 6$ and $k = 6$,

(ii) $\ell > 7$ and $k = 7$,

(iii) $\ell < 9$ and $k = 9$, or

(iv) $\ell \leq 7$, $s'_7[5] = 4$ and $k = 8$.

Proof There is no new permissible function on the lingering loop γ_ℓ. Suppose $k \neq \ell$. Let j be the unique integer satisfying $s'_k[j] = s_k[j] + 1$. There is a new permissible function in $\mathcal{B}$ on γ_k if and only if either the function φ_{jj} is both new and departing, or there is an integer i such that $s'_k(\varphi_{ij}) = s_k(\theta)$. We now examine when such an i exists.

The values $s'_k[i]$ are six distinct integers between -2 and 5. Let a and b be the two integers in this range that are not equal to $s'_k[i]$ for any i. On the hth nonlingering loop, one has

$$h = \sum_{i=0}^{5}(s'_k[i] + 2 - i) = 9 - (a + b).$$

Since $s'_k[j] = s_k[j] + 1$, we must have that $s'_k[j]$ is equal to either $a + 1$ or $b + 1$. Without loss of generality, assume that it is equal to $a + 1$. There does not exist i such that $s'_k[i] + s'_k[j] = s'_k(\theta)$ if and only if $s'_k(\theta) - (a + 1)$ is greater than 5, smaller than -2, or equal to either a or b. If it is equal to a, then

the function φ_{jj} is both new and departing. Since $s'_k(\theta) \leq 4$ and $a + 1 \geq -1$, we see that $s'_k(\theta) - (a + 1)$ cannot be greater than 5, and $s'_k(\theta) - (a + 1)$ is smaller than -2 if and only if $s'_k(\theta) = 2$ and $a = 4$. By the above calculation, $b = s'_k(\theta) - (a + 1)$ if and only if $h = 10 - s'_k(\theta)$.

The 6th nonlingering loop is contained in the first block if and only if $\ell > 6$. The 7th nonlingering loop is contained in the second block if and only if $\ell > 7$. The 8th nonlingering loop is contained in the third block if and only if $\ell < 9$. Finally, if $a = 4$, then γ_k is one of the first 7 nonlingering loops. If γ_k is in the third block, then since $z_2 \geq 7$, we have $\ell \leq 7$, and γ_k is the first loop in the third block. $\square$

Having determined which loops have new permissible functions in $\mathcal{B}$, we can now strategically choose the subset $\mathcal{B}' \subset \mathcal{B}$ from which we will construct the independence θ, so that the number of permissible functions in $\mathcal{B}'$ on each block is precisely one more than the number of loops in the block. Note that $|\mathcal{B}| = 21$, so $\mathcal{B}'$ is chosen by omitting a single function ψ from $\mathcal{B}$.

Definition 4.8 If $\ell \leq 7$, let $\psi \in \mathcal{B}$ be a function that is permissible on the second block. Otherwise, let $\psi \in \mathcal{B}$ be a function that is permissible on the third block. Let $\mathcal{B}' = \mathcal{B} \setminus \{\psi\}$.

Remark 4.9 There may be several functions that are permissible on the specified block; it does not matter which of these we omit from $\mathcal{B}'$.

Lemma 4.10 *On each block, the number of permissible functions in $\mathcal{B}'$ is one more than the number of loops.*

Proof This follows directly from Proposition 4.7. Specifically, since $z_1 = \min\{6, \ell\}$, there is a new permissible function in $\mathcal{B}$ on each loop of the first block, except for the last one. Since there are precisely two pairs (i, j) such that $s_1(\varphi_{ij}) = 4$, we see that the number of permissible functions on the first block is one more than the number of loops. By symmetry, if $z_2 \leq 7$, then the number of permissible functions in $\mathcal{B}$ on the third block is one more than the number of loops, and if $z_2 > 7$, it is two more. But when $z_2 > 7$, one of these functions is not in $\mathcal{B}'$.

Finally, we consider the middle block. We count the number of pairs (i, j) such that $s'_{z_1}(\varphi_{ij}) = 3$. Since 3 is odd, if (i, j) is such a pair, then $i \neq j$. It follows that there are 3 such pairs if and only if $s'_{z_1}[i] + s'_{z_1}[5 - i] = 3$ for all i, which implies that there are precisely 6 nonlingering loops in the first block. It follows that, if $\ell < 7$, then there are precisely two such pairs, and if $\ell \geq 7$, there are three such pairs. By Proposition 4.7, if $\ell < 7$, there is a new permissible function on every loop of the middle block. If $\ell = 7$, then the middle block contains only one loop, and since this loop is lingering, there are no new permissible functions on it. In both of these cases, the number of permissible functions in $\mathcal{B}$ on the middle block is therefore two more than the number of loops, but one of these functions is not in $\mathcal{B}'$. If $\ell > 7$, then there are no new permissible functions on γ_7 or γ_ℓ, so the number of permissible functions is one more than the number of loops. $\square$

We now describe the algorithm for constructing our independence

$$\theta = \min_{\varphi_{ij} \in \mathcal{B}'} \{\varphi_{ij} + c_{ij}\},$$

with slopes $s_k(\theta)$ as specified in (21), when $g = 13$, D is vertex-avoiding, and Σ is unramified. The algorithm is quite similar to that presented in [19, Section 8.1]. We include the details. See Example 4.19 for an illustration of the output in one particular case.

In this algorithm, we move from left to right across each of the three blocks where $s_k(\theta)$ is constant, adjusting the coefficients of unassigned permissible functions and assigning one function $\varphi_{ij} \in \mathcal{B}'$ to each loop so that each function achieves the minimum uniquely on some part of the loop to which it is assigned. At the end of each block, there is one remaining unassigned permissible function that achieves the minimum uniquely on the bridge immediately after the block, which we assign to that bridge. Since there are 13 loops and three blocks, this gives us an independent configuration of 16 functions. The remaining 4 functions, with slopes too high or too low to be permissible on any block, achieve the minimum uniquely on the bridges to the left of the first loop or to the right of the last loop, respectively. Example 4.19 illustrates the procedure for one randomly chosen tableau. We now list a few of the key properties of the algorithm:

(i)   Once a function has been assigned to a bridge or loop, it always achieves the minimum uniquely at some point on that bridge or loop.

(ii)   A function never achieves the minimum on any loop to the right of the bridge or loop to which it is assigned.

(iii)   The coefficient of each function is initialized to ∞ and then assigned a finite value when the function is assigned to a bridge or becomes permissible on a loop, whichever comes first.

(iv)   After the initial assignment of a finite coefficient, subsequent adjustments to this coefficient are smaller and smaller perturbations. This is related to the fact that the edges get shorter and shorter as we move from left to right across the graph.

(v)   Only the coefficients of unassigned functions are adjusted, and all adjustments are upward. This ensures that once a function is assigned and achieves the minimum uniquely on a loop, it always achieves the minimum uniquely on that loop.

(vi)   Exactly one function is assigned to each of the 13 loops, and the remaining seven functions are assigned to either the leftmost bridge, the rightmost bridge, or one of the three bridges after the blocks.

The algorithm terminates when we reach the rightmost bridge, at which point each of the 20 functions $\{\varphi_{ij} + c_{ij} : \varphi_{ij} \in \mathcal{B}'\}$ achieves the minimum uniquely at some point on the graph.

Remark 4.11   The one crucial difference, in comparison with the construction in [19, Section 8.1], is that we do *not* skip the lingering loop γ_ℓ. Instead, since γ_ℓ is the last loop in its block, there are precisely

two unassigned permissible functions on γ_ℓ. These two functions do not have identical restrictions to γ_ℓ. Thus, if we adjust their coefficients upward so that they agree at w_ℓ, one of them will obtain the minimum uniquely at some point of the loop γ_ℓ. We assign this function to γ_ℓ and adjust its coefficient upward by an amount small enough so that it still obtains the minimum uniquely at some point of γ_ℓ. The other achieves the minimum uniquely at w_ℓ, and we assign it to the bridge $\beta_{\ell+1}$.

The algorithm depends on the following basic properties of the permissible functions φ_{ij}.

Lemma 4.12 *There is at most one departing permissible function φ_{ij} on each loop γ_k. Furthermore, if γ_k is lingering then there are none.*

Proof The proof is identical to [19, Lemma 8.8]. □

Lemma 4.13 *For any loop γ_k, there are at most three nondeparting permissible functions in $\mathcal{B}$ on γ_k.*

Proof If φ_{ij} is a nondeparting permissible function on γ_k, then $s_{k+1}(\varphi_{ij}) = s_k(\theta)$. For each i, this equality holds for at most one j, and the lemma follows. □

Proposition 4.14 *Consider a set of at most three nondeparting permissible functions from $\mathcal{B}$ on a loop γ_k and assume that all of the functions take the same value at w_k. Then there is a point of γ_k at which one of these functions is strictly less than the others.*

Proof The proof is identical to [19, Lemma 8.19]. □

The algorithm is as follows:

- **Start at the first bridge** Start at β_1 and initialize $c_{55} = 0$. Initialize c_{45} so that $\varphi_{45} + c_{45}$ equals φ_{55} at a point one third of the way from w_0 to v_1. Initialize c_{44} and c_{35} so that $\varphi_{44} + c_{44}$ and $\varphi_{35} + c_{35}$ agree with $\varphi_{45} + c_{45}$ at a point two thirds of the way from w_0 to v_1. Initialize all other coefficients c_{ij} to ∞. Note that φ_{55} and φ_{45} achieve the minimum uniquely on the first and second third of β_1, respectively. Assign both of these functions to β_1, and proceed to the first loop.

- **Loop subroutine** Each time we arrive at a loop γ_k, apply the following steps:

  - **Step 1: reinitialize unassigned coefficients** By Lemma 4.15 below, there are at least two unassigned permissible functions. Find the unassigned permissible function φ_{ij} that maximizes $\varphi_{ij}(w_k) + c_{ij}$. Initialize the coefficients of the new permissible functions (if any) and adjust the coefficients of the other unassigned permissible functions upward so that they all agree with φ_{ij} at w_k. (The unassigned permissible functions are strictly less than all other functions on γ_k, even after this upward adjustment; see Lemma 4.16.)

  - **Step 2: assign departing functions** If there is a departing function, assign it to the loop. (There is at most one, by Lemma 4.12.) Adjust the coefficients of the other permissible functions upward

so that all of the functions agree at a point on the following bridge a short distance to the right of w_k, but far enough so that the departing function achieves the minimum uniquely on the whole loop. This is possible because the bridge is much longer than the edges in the loop. Proceed to the next loop.

- **Step 3: otherwise, use Proposition 4.14**   By Lemma 4.13, there are at most three nondeparting functions. By Proposition 4.14, there is one φ_{ij} that achieves the minimum uniquely at some point of γ_k. We adjust the coefficient of φ_{ij} upward by $\frac{1}{3}m_k$. This ensures that it will never achieve the minimum on any loops to the right, yet still achieves the minimum uniquely on this loop; see Lemma 4.16, below. Assign φ_{ij} to γ_k, and proceed to the next loop.

- **Proceeding to the next loop**   If the next loop is contained in the same block, then move right to the next loop and apply the loop subroutine. Otherwise, the current loop is the last loop in its block. In this case, proceed to the next block.

- **Proceeding to the next block**   After applying the loop subroutine to the last loop in a block, there is exactly one unassigned permissible function in $\mathcal{B}'$, by Lemma 4.10. The unassigned permissible function φ_{ij} achieves the minimum uniquely on the beginning of the outgoing bridge, without any further adjustment of coefficients. Assign φ_{ij} to this bridge.

If we are at the last loop γ_g, then proceed to the last bridge. Otherwise, there are several permissible functions on the first loop of the next block, as detailed in Lemma 4.13, above. Initialize the coefficient of each permissible function on the first loop of the next block so that it is equal to θ at the midpoint of the bridge between the blocks, and then apply the loop subroutine.

- **The last bridge**   Initialize the coefficient c_{01} so that $\varphi_{01} + c_{01}$ equals θ at the midpoint of the last bridge β_{14}. Initialize c_{00} so that $\varphi_{00} + c_{00}$ equals θ halfway between the midpoint and the rightmost endpoint. Note that both of these functions now achieve the minimum uniquely at some point on the second half of β_{14}. Assign both of these functions to β_{14}, and output $\theta = \min\{\varphi_{ij} + c_{ij} : \varphi_{ij} \in \mathcal{B}'\}$.

To verify that this algorithm produces a tropical independence, we first show that there are at least two unassigned permissible functions on each loop.

Lemma 4.15   *There are at least two unassigned permissible functions on each loop γ_k.*

Proof   By Lemma 4.10, the number of permissible functions in $\mathcal{B}'$ on the block containing γ_k is one more than the number of loops. Since there is at most one new function per loop, the number of functions in $\mathcal{B}'$ that are permissible on some loop between the first loop of the block and γ_k, inclusive, is at least one more than the number of loops. Finally, note that exactly one function is assigned to each loop, and moreover, if a function is departing, it is assigned. It follows by induction on k' that the number of functions in $\mathcal{B}'$ that are unassigned and permissible on some loop between $\gamma_{k'}$ and γ_k is at least $k - k' + 2$. Hence, the number of unassigned permissible functions on γ_k is at least two.   □

We now verify that this algorithm produces a tropical independence.

Lemma 4.16 *Suppose that φ_{ij} is assigned to the loop γ_k or the bridge β_k. Then φ_{ij} does not achieve the minimum at any point to the right of v_{k+1}.*

Proof If γ_k is a nonlingering loop, then the proof is the same as [19, Section 8.2]. On the other hand, if γ_k is the lingering loop, then it is the last loop in its block. Since v_{k+1} is the start of the next block, φ_{ij} cannot achieve the minimum at any point to the right of v_{k+1}. $\qquad\square$

This completes the proof of Theorem 4.2 in the vertex-avoiding case.

Remark 4.17 For future reference, we note that the proof of Lemma 4.16 does not depend on the relative lengths of the bridges. It only uses that the bridges are much longer than the loops. The assumption that each bridge is much longer than the next is only used later, when there are decreasing bridges, decreasing loops, or switching loops.

Remark 4.18 If Γ' is the subgraph of Γ to the right of w_1, then Γ' is a chain of 12 loops whose edge lengths satisfy the required conditions, and if the first loop is nonlingering, then the restriction of Σ to Γ' satisfies the ramification condition of Theorem 4.2, with equality. Similarly, the subgraph to the right of w_2 is a chain of 11 loops whose edge lengths satisfy the required conditions, and the restriction of Σ to this subgraph satisfies the ramification condition of Theorem 4.2, with equality. To produce an independence in these cases, assign each function in $\mathcal{B}'$ with slope greater than 4 to the first bridge, and then proceed as above. There are precisely $15 - g$ such functions, and they have distinct slopes along the first bridge, as in [19, Lemma 10.40]. Because of this, we can choose coefficients so that each one obtains the minimum uniquely at some point of the first bridge. Thus the unramified vertex-avoiding cases of Theorem 4.2 for $g = 11$ and 12 (ie when Σ is unramified at v_{14} and there is no extra ramification at w_{13-g} beyond what is required by the inequalities on vanishing orders in the statement of the theorem) follow from essentially the same argument as for $g = 13$. Our choice to index the vertices starting at w_{13-g} reflects the idea that these linear series with ramification on a chain of $g = 11$ or 12 loops behave like linear series on a chain of 13 loops restricted to the subgraph to the right of w_{13-g}.

Example 4.19 We illustrate the construction with an example. Let $[D]$ be a vertex-avoiding class of degree 16 and rank 5 associated to the tableau in Figure 3.

The independence $\theta = \min_{ij}\{\varphi_{ij} + c_{ij}\}$ that we construct is depicted schematically in Figure 4. The graph should be read from left to right and top to bottom, so the first six loops appear in the first row, with

1	3	4	8	9	10
2	5	7	11	12	13

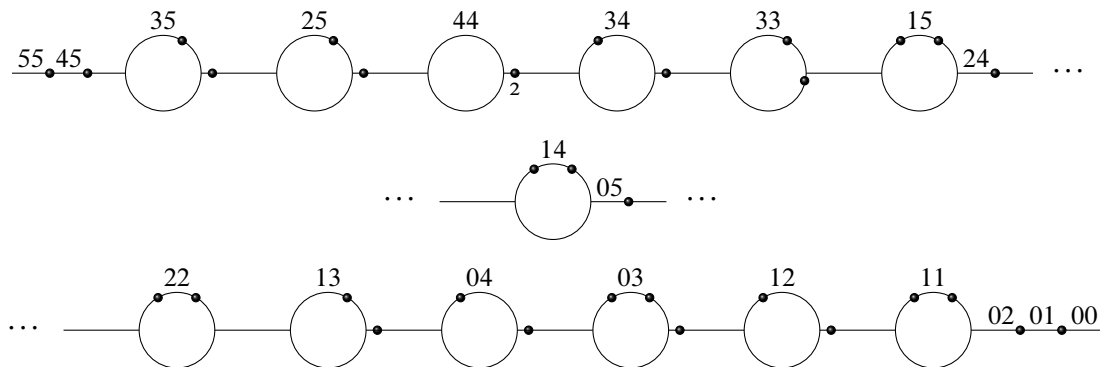Figure 3: The tableau corresponding to the divisor D.

Figure 4: The divisor $D' = 2D + \operatorname{div}(\theta)$. The function φ_{ij} achieves the minimum uniquely on the region labeled ij in $\Gamma \setminus \operatorname{Supp}(D')$.

γ_1 on the left and γ_6 on the right, and γ_{13} is the last loop in the third row. The rows correspond to the three blocks. The 31 dots indicate the support of the divisor $D' = 2D + \operatorname{div}(\theta)$. Note that $\deg(D') = 32$; the point on the bridge β_4 appears with multiplicity 2, as marked. Because $\ell = 6$, there is a function that is permissible on the second block in $\mathcal{B}$ but not $\mathcal{B}'$. The functions in $\mathcal{B}$ that are permissible on the second block are precisely φ_{05}, φ_{14}, and φ_{23}; we have chosen (arbitrarily) to omit φ_{23} from $\mathcal{B}'$. Each of the 20 functions φ_{ij} in $\mathcal{B}'$ achieves the minimum uniquely on the connected component of the complement of $\operatorname{Supp}(D')$ labeled ij.

4.2 No switching loops

Recall that a loop γ_ℓ is a *switching loop* for Σ if there is some $\varphi \in \Sigma$ and some h such that $s_\ell(\varphi) \leq s_\ell[h]$ and $s'_\ell(\varphi) \geq s'_\ell[h+1]$. It is a *lingering loop* if it is not a switching loop and $s_\ell[i] = s'_\ell[i]$ for all i. Recall also that γ_ℓ is a *decreasing loop* if $s_\ell[h] > s'_\ell[h]$. Similarly β_ℓ is a *decreasing bridge* if $s'_{\ell-1}[h] > s_\ell[h]$.

Because we are only considering cases where the adjusted Brill–Noether number is at most one, by [19, Proposition 9.10], we know that there is at most one lingering loop, one positive ramification weight, one decreasing loop, one decreasing bridge, or one switching loop, and these possibilities are mutually exclusive. Moreover, for decreasing loops and bridges, the index h is unique and the decrease in slope is exactly one. In this subsection, we consider all cases where there is no switching loop. The cases with a switching loop are discussed in Section 4.3.

Assume Σ has no switching loops. Then for all i there is a function $\varphi_i \in \Sigma$ such that

$$s_k(\varphi_i) = s_k[i] \quad \text{and} \quad s'_k(\varphi_i) = s'_k[i] \quad \text{for all } k.$$

We keep the notation $\varphi_{ij} = \varphi_i + \varphi_j$ and $\mathcal{B} = \{\varphi_{ij} : 0 \leq i \leq j \leq 5\}$. As in the unramified vertex-avoiding case, we choose a subset $\mathcal{B}' \subseteq \mathcal{B}$ of 20 functions, and we choose integers z_1 and z_2 in order to divide the graph Γ into three blocks. We make our choices to satisfy the following conditions:

(i) No two functions in $\mathcal{B}'$ that are permissible on γ_k differ by a constant on γ_k.

(ii)   The number of functions in $\mathcal{B}'$ that are permissible on each block is at most one more than the number of loops in that block.

(iii)   No function in $\mathcal{B}'$ is permissible on more than one block,

(iv)   if γ_k is a lingering loop, then it is the last loop in its block.

(v)   If γ_k is a decreasing loop and j is the unique value such that $s'_k[j] < s_k[j]$, then no function of the form $\varphi_{ij} \in \mathcal{B}'$ is permissible on γ_k.

(vi)   If β_k is a decreasing bridge and j is the unique value such that $s_k[j] < s'_{k-1}[j]$, then either β_k is a bridge between blocks, or no function of the form $\varphi_{ij} \in \mathcal{B}'$ is permissible on γ_{k-1}.

Proposition 4.20   *If $\mathcal{B}'$ satisfies conditions* (i)–(vi), *then the functions in $\mathcal{B}'$ are independent.*

Proof   The algorithm for constructing the tropical independence is identical to the algorithm of Section 4.1, with the following exceptions. First, as in Remark 4.18, we assign every function with slope greater than four to the first bridge. Second, the procedure for proceeding to the next block must be altered slightly when the bridge between the blocks is a decreasing bridge.

When the bridge between the blocks is a decreasing bridge, there is a unique point v on the bridge where one of the functions φ_i is locally nonlinear. We initialize the coefficients of the new permissible functions on the next block so that they are equal to θ at a point to the right of v. If one of the blocks contains zero loops, we set the coefficient of the unique function with slope equal to that of θ so that it is equal to θ at a point to the right of v, and initialize the coefficients of the new permissible functions on the next block so that they are equal to θ at a point to the right of this.

We note that there are at most 3 nondeparting permissible functions in $\mathcal{B}'$ on each loop. This is because a nondeparting permissible function φ_{ij} on γ_k satisfies $s_{k+1}(\varphi_{ij}) = s_k(\theta)$, and for each i this equality can hold for at most one j.

To see that this algorithm produces an independence, suppose that φ_{ij} is assigned to the loop γ_k or the bridge β_k. We show that φ_{ij} does not achieve the minimum at any point to the right of v_{k+1}. If γ_k and β_k both have multiplicity zero, then the argument is the same as in [19, Section 8.2]. On the other hand, if γ_k has positive multiplicity, then either γ_k is a decreasing loop, or by (iv) it is the last loop in its block. If γ_k is a decreasing loop, then by (v) there is no function in $\mathcal{B}'$ that is permissible on γ_k and contains the decreasing function as a summand, so the result holds again as in [19, Section 8.2]. We may therefore assume that γ_k is the last loop in its block, in which case the argument is identical to the vertex-avoiding case above.

Similarly, if β_k has positive multiplicity, then by (vi) there are two possibilities. If φ_{ij} does not contain the decreasing function as a summand, then there is nothing to show. Otherwise, β_k is a bridge between blocks. By (iii) the function φ_{ij} is only permissible on one block. Since v_{k+1} is the start of the next block, φ_{ij} cannot achieve the minimum at any point to the right of v_{k+1}.   □

For the rest of this section, we explain how to choose z_1, z_2, and the set $\mathcal{B}'$ in order to satisfy conditions (i)–(vi). This is done by a careful case analysis, depending on combinatorial properties of the tropical linear series Σ.

Case 1: there are no loops or bridges of positive multiplicity This guarantees that either the linear series is ramified at v_{14}, or has "extra ramification" at w_{13-g}, meaning that $g = 13$ and the linear series is ramified at w_0, or $g = 11$ or 12 and the linear series has more ramification than what is imposed by the inequalities on vanishing numbers in Theorem 4.2s. In these cases, which are mutually exclusive, we choose z_1 and z_2 so that γ_{z_1} is the first loop in the first block with no new function, and γ_{z_2+1} is the last loop in the last block with no departing function. These loops exist by a counting argument, but we make the choice explicit.

If Σ is ramified at v_{14}, let k be the smallest positive integer such that $s'_k[5] = 6$, and define

(22) $$z_1 = \begin{cases} 6 & \text{if } k \geq 7, \\ 7 & \text{if } k \leq 6, \end{cases} \quad \text{and} \quad z_2 = \max\{k-1, 7\}.$$

If Σ has extra ramification at w_{13-g}, let k be the largest positive integer such that $s_k[0] = -3$, and define

(23) $$z_1 = \min\{k, 6\} \quad \text{and} \quad z_2 = \begin{cases} 6 & \text{if } k \geq 8, \\ 7 & \text{if } k \leq 7. \end{cases}$$

Let $\psi \in \mathcal{B}$ be a function that is permissible on the second block, and let $\mathcal{B}' = \mathcal{B} \setminus \{\psi\}$. (In the case where $z_1 = z_2$, let $\psi \in \mathcal{B}$ be a function with $s_{z_1+1}(\psi) = 3$.)

If there is a loop or bridge of positive multiplicity, then since $\rho = 1$, there is only one such loop or bridge, and it has multiplicity 1.

Case 2: there is a bridge β_ℓ of multiplicity 1 If $\ell \geq 8$ and $s'_{\ell-1}[5] = 6$, then define z_1 and z_2 as in (22). If $\ell \leq 7$ and $s_\ell[0] = -3$, then define z_1 and z_2 as in (23). Otherwise, define

$$z_1 = \min\{\ell-1, 6\} \quad \text{and} \quad z_2 = \ell - 1.$$

If $\ell \geq 8$ and $s_{\ell-1}[5] = 6$, or $\ell \leq 7$ and $s_\ell[0] = -3$, then as above, we let $\psi \in \mathcal{B}$ be a function that is permissible on the second block, and let $\mathcal{B}' = \mathcal{B} \setminus \{\psi\}$. Otherwise, let h be the unique integer such that $s_\ell[h] < s'_{\ell-1}[h]$. If $\ell \neq 5, 6$, then we will see in Lemma 4.21 that either there is a unique i such that $s'_{\ell-1}[h] + s'_{\ell-1}[i] = s_{\ell-1}(\theta)$, or $2s'_{\ell-1}[h] = s_{\ell-1}(\theta) + 1$, but not both. In the first case, we let $\mathcal{B}' = \mathcal{B} \setminus \{\varphi_{hi}\}$, and in the second case, we let $\mathcal{B}' = \mathcal{B} \setminus \{\varphi_{hh}\}$. (The function in $\mathcal{B} \setminus \mathcal{B}'$ is permissible on both blocks to either side of the bridge β_ℓ.) If $\ell = 5$ or 6, then we will see in Lemma 4.21 that there is a unique i such that $s'_{\ell-1}[h] + s'_{\ell-1}[i] = s_{\ell-1}(\theta) - 1$, and we again let $\mathcal{B}' = \mathcal{B} \setminus \{\varphi_{hi}\}$.

It remains to consider the cases where there is a loop of multiplicity one. The case of a switching loop is left to the next subsection. In the case of a lingering loop, we construct an independence exactly as in Section 4.1. (See Remark 4.18 for an explanation of how the algorithm for $g = 13$ is adapted to the cases where $g = 11$ or $g = 12$.) We now discuss the remaining case, where there is a decreasing loop.

Case 3: there is a decreasing loop γ_ℓ If $\ell \geq 8$ and $s_\ell[5] = 6$, then define z_1 and z_2 as in (22). If $\ell \leq 7$ and $s'_\ell[0] = -3$, then define z_1 and z_2 as in (23). Otherwise, define

$$z_1 = \begin{cases} \ell & \text{if } \ell < 6, \\ 5 & \text{if } \ell = 6, \\ 6 & \text{if } \ell > 6, \end{cases} \quad \text{and} \quad z_2 = \begin{cases} \ell - 1 & \text{if } \ell > 8, \\ 8 & \text{if } \ell = 8, \\ 7 & \text{if } \ell < 8. \end{cases}$$

If $\ell \geq 8$ and $s_\ell[5] = 6$ or $\ell \leq 7$ and $s_\ell[0] = -3$, then as above, we let $\psi \in \mathcal{B}$ be a function that is permissible on the second block, and let $\mathcal{B}' = \mathcal{B} \setminus \{\psi\}$. Otherwise, let h be the unique integer such that $s'_\ell[h] < s_\ell[h]$. If $\ell < 6$ or $\ell = 7, 8$ then γ_ℓ is the last loop in its block, and we will see in Lemma 4.21 that either there is a unique i such that $s_\ell[h] + s_\ell[i] = s_\ell(\theta)$, or $2s_\ell[h] = s_\ell(\theta) + 1$, but not both. In the first case, we let $\mathcal{B}' = \mathcal{B} \setminus \{\varphi_{hi}\}$, and in the second case, we let $\mathcal{B}' = \mathcal{B} \setminus \{\varphi_{hh}\}$. If $\ell > 8$ or $\ell = 6$, then we will see that either there is a unique i such that $s_\ell[h] + s_\ell[i] = s_{\ell-1}(\theta)$, or $2s_\ell[h] = s_{\ell-1}(\theta) + 1$. Again, in the first case, we let $\mathcal{B}' = \mathcal{B} \setminus \{\varphi_{hi}\}$, and in the second case, we let $\mathcal{B}' = \mathcal{B} \setminus \{\varphi_{hh}\}$.

In the cases above, we asserted several times that certain functions exist with specified slopes. To prove this, we need to generalize Proposition 4.7. We first define the function

$$\tau(k) = \sum_{i=0}^{5} (s'_k[i] + 2 - i).$$

Note that, if there is a loop of positive multiplicity and γ_ℓ is the kth loop of multiplicity zero, then $k = \tau(\ell)$. The following observation serves as the basis for our counting arguments.

Lemma 4.21 *For a fixed k, suppose that $-2 \leq s'_k[i] \leq 5$ for all i. Let j be an integer such that $s'_k[j] - 1$ is not equal to -3 or $s'_k[i]$ for any i. For s in the range $2 \leq s \leq 5$, there does not exist i such that $s'_k[i] + s'_k[j] = s$ if and only if one of the following holds:*

(i) $\tau(k) = 10 - s$.

(ii) $s = 5$, $j = 0$ and $s'_k[0] = -1$.

(iii) $s = 2$, $j = 5$ and $s'_k[5] = 5$.

(iv) $2s'_k[j] = s + 1$.

Proof The argument is identical to that of Proposition 4.7. □

There are additional relevant cases, when $s'_k[5] = 6$ or $s'_k[0] = -3$.

Lemma 4.22 *If $s'_k[5] = 6$, then there does not exist i such that $s'_k[i] + 6 \leq 3$. Similarly, if $s_k[0] = -3$, then there does not exist i such that $s'_k[i] - 2 \geq 4$.*

Proof Since $\rho = 1$, if $s'_k[5] = 6$, then $s'_k[0] \geq -2$. It follows that $s'_k[i] + 6 \geq 4$ for all i. Similarly, if $s_k[0] = -3$, then $s'_k[i] \leq 5$ for all i. It follows that $s'_k[i] - 2 \leq 3$ for all i. □

Lemma 4.23 *The set $\mathcal{B}'$ satisfies conditions* (i)–(vi).

Proof (i) If $s'_k[i] \geq s_k[i]$ for all i, then the result is immediate, so we may assume that γ_k is a decreasing loop. Let h be the unique integer such that $s'_k[h] = s_k[h] + 1$, and let h' be the unique integer such that $s'_k[h'] = s_k[h'] - 1$. If $\varphi_{hh'}$ is not permissible on γ_k, then again there is nothing to show. If $\varphi_{hh'}$ is permissible, then by Lemma 4.21, we must have $s_k(\theta) = 10 - k$. By construction, this occurs if and only if $k = 7$, in which case $\varphi_{hh'} \notin \mathcal{B}'$.

(ii) Consider the first block first. There are two functions $\psi \in \mathcal{B}$ with the property that $s'_{13-g}(\psi) = 4$. The result will therefore hold for the first block if and only if the first block contains a loop with no new permissible functions. Let γ_k be a loop of multiplicity zero that is contained in the first block. By Lemmas 4.21 and 4.22, there is no new permissible function on γ_k if and only if $\tau(k) = 6$ or $s'_k[0] = s_k[0] + 1 = -2$. Thus, the number of permissible functions in $\mathcal{B}$ on the first block is at most two more than the number of loops in Cases 2 or 3 when $\ell \leq 6$ and $s_\ell[0] \geq -2$, and one more than the number of loops in the remaining cases. In Cases 2 and 3 when $\ell \leq 6$ and $s_\ell[0] \geq -2$, the function in $\mathcal{B} \setminus \mathcal{B}'$ is permissible on the first block. Since this function is not in $\mathcal{B}'$, the number of functions in $\mathcal{B}'$ that are permissible on the first block is one less than the number in $\mathcal{B}$. The third block follows from a completely symmetric argument.

For the second block, note that if $\tau(z_1) = 6$, then there are 3 functions $\psi \in \mathcal{B}$ with the property that $s'_{z_1}(\psi) = 3$, and otherwise there are only two such functions. In every case, either $\tau(z_1) < 6$ or by Lemma 4.21, the second block contains a loop with no new permissible functions. Since the function in $\mathcal{B} \setminus \mathcal{B}'$ is permissible on the second block, we see that the number of permissible functions on the second block is one more than the number of loops. (Note that this holds even in the case where the second block contains zero loops, in which case there is exactly one permissible function on the second block.)

(iii) Suppose that $\varphi_{ij} \in \mathcal{B}$ is permissible on more than one block. First, consider the case where β_ℓ is a bridge of multiplicity one, and let h be the unique integer such that $s_\ell[h] = s_{\ell-1}[h] - 1$. If φ_{ij} is permissible on more than one block, then $j = h$ and either $s'_{\ell-1}[h] + s'_{\ell-1}[i] = s_{\ell-1}(\theta)$, or $i = h$ and $2s'_{\ell-1}[h] = s_{\ell-1}(\theta) + 1$. If $-2 \leq s_\ell[h] \leq 5$, then by Lemma 4.21, such an i exists if and only if $\ell \neq 5, 6$, and by construction, we have $\varphi_{hi} \notin \mathcal{B}'$. Similarly, if $s_\ell[h] = -3$, then by Lemma 4.22, such an i exists if and only if $\ell \geq 8$, and if $s_\ell[h] = 5$, then such an i exists if and only if $\ell \leq 7$. In both cases, we have $\varphi_{hi} \notin \mathcal{B}'$.

Next, consider the case where γ_ℓ is a decreasing loop. By construction, γ_ℓ is either the first or last loop in its block. Let h be the unique integer such that $s'_\ell[h] = s_\ell[h] - 1$. If γ_ℓ is the last loop in its block and φ_{ij} is permissible on both the block containing γ_ℓ and the next block, then $j = h$ and either $s_\ell[h] + s_\ell[i] = s_\ell(\theta)$, or $i = h$ and $2s_\ell[h] = s_\ell(\theta) + 1$. But then $\varphi_{ij} \notin \mathcal{B}'$. Similarly, if γ_ℓ is the first loop in its block, and φ_{ij} is permissible on both the block containing γ_ℓ and the preceding block, then $j = h$ and either $s_\ell[h] + s_\ell[i] = s_{\ell-1}(\theta)$, or $2s_\ell[h] = s_{\ell-1}(\theta) + 1$. If $\ell \neq 7$, then again $\varphi_{ij} \notin \mathcal{B}'$. Finally,

note that if γ_ℓ is *both* the first and last loop in its block, then $\ell = 7$, and the only functions φ_{ij} that are permissible on γ_7 satisfy $s'_\ell[i] + s'_\ell[j] = 3$. The result follows.

(iv) If γ_ℓ is a lingering loop, then we follow the construction of the vertex-avoiding case of the previous subsection, in which γ_ℓ is the last loop in its block.

(v) Let γ_k be a decreasing loop, let h be the unique integer such that $s'_k[h] = s_k[h] + 1$, and let h' be the unique integer such that $s'_k[h'] = s_k[h'] - 1$. If $\varphi_{hh'}$ is permissible, then $\varphi_{hh'} \notin \mathcal{B}'$, as shown in the proof of condition (i).

(vi) Let β_k be a decreasing bridge and let j is the unique value such that $s_k[j] < s'_{k-1}[j]$. If β_k is not a bridge between blocks, then by construction either $j = 0$, $k \le 7$, and $s_k[0] = -3$, or $j = 5$, $k \ge 8$, and $s_k[5] = 5$. In both cases, by Lemma 4.22, we see that there is no i such that $\varphi_{ij} \in \mathcal{B}'$ is permissible on γ_{k-1}. $\qquad\square$

This completes the proof of Theorem 4.2 in all cases where there is no switching loop for Σ.

4.3 Switching loops

We now consider the case where there is a switching loop γ_ℓ that switches slope h. This means that $s_\ell[i] = s'_\ell[i]$ for all i, and there exists a function $\varphi \in R(D)$ satisfying

$$s_\ell(\varphi) = s_\ell[h] \quad \text{and} \quad s'_\ell(\varphi) = s'_\ell[h] + 1 = s'_\ell[h+1].$$

In this case, we define z_1 and z_2 as follows:

$$z_1 = \begin{cases} \ell & \text{if } \ell < 6, \\ 5 & \text{if } \ell = 6, \\ 6 & \text{if } \ell > 6, \end{cases} \quad \text{and} \quad z_2 = \begin{cases} 7 & \text{if } \ell < 6, \\ \ell & \text{if } \ell \ge 6. \end{cases}$$

As in Section 4.1, we will construct our independence θ to satisfy

$$s_k(\theta) = \begin{cases} 4 & \text{if } k \le z_1, \\ 3 & \text{if } z_1 < k \le z_2, \\ 2 & \text{if } z_2 < k \le 13. \end{cases}$$

In the preceding cases, we identified functions $\varphi_i \in \Sigma$ with designated slope $s_k(\varphi_i) = s_k[i]$ along each bridge β_k. When there is a switching loop, this is possible for $i \ne h, h+1$, but such a function does not necessarily exist for $i = h, h+1$. Instead, we identify a collection of functions in Σ with designated slope along some of the bridges, and with slopes along the remaining bridges in a restricted range.

Proposition 4.24 *There is a pencil $W \subseteq V$ with φ_A, φ_B and φ_C in $\mathrm{trop}(W)$ such that*

(i) $s'_k(\varphi_A) = s'_k[h]$ *for all* $k < \ell$,

(ii) $s_k(\varphi_B) = s_k[h+1]$ *for all* $k > \ell$,

(iii) $s_k(\varphi_C) = s_k[h+1]$ *for all* $k \le \ell$, *and* $s'_k(\varphi_C) = s_k[h]$ *for all* $k \ge \ell$,

(iv) $s_k(\varphi_\bullet) \in \{s_k[h], s_k[h+1]\}$ *and* $s'_k(\varphi_\bullet) \in \{s'_k[h], s'_k[h+1]\}$ *for all* k.

Proof The proof is essentially the same as that of [19, Proposition 11.18]. We include the details for completeness. First, there exists $\varphi_A \in \Sigma$ such that $s'_{13-g}(\varphi_A) \leq s'_{13-g}[h]$ and $s_{14}(\varphi_A) \geq s_{14}[h]$. Since γ_ℓ is the only switching loop, we have $s'_k(\varphi_A) \leq s'_k[h]$ for $k < \ell$, and $s'_k(\varphi_A) \geq s'_k[h]$ for $k \geq \ell$. In particular, $s'_\ell(\varphi_A) \geq s'_\ell[h]$, so $s'_{\ell-1}(\varphi_A) \geq s'_{\ell-1}[h]$, and it follows that $s'_{\ell-1}(\varphi_A) = s'_{\ell-1}[h]$. This proves (i), because there are no switching loops to the left of γ_ℓ. The proof of (ii) is similar.

We now prove (iii). Given φ_A and φ_B in Σ satisfying (i) and (ii), choose f_A and $f_B \in V$ tropicalizing to φ_A and φ_B, respectively. Let W be the pencil spanned by f_A and f_B. Arguments similar to the proof of (i) above show that $s_k(\mathrm{trop}(W)) = (s_k[h], s_k[h+1])$, for all k. Choose a function $f \in W$ such that $\varphi = \mathrm{trop}(f)$ satisfies $s_\ell(\varphi) = s_\ell[h+1]$. Then $s_k(\varphi) = s_k[h+1]$ for $k < \ell$. Similarly, choose $\varphi' \in \mathrm{trop}(W)$ such that $s'_\ell(\varphi') = s'_\ell[h]$, which implies that $s_k(\varphi') = s_k[h]$ for $k > \ell$. Finally, by adding a scalar to φ', we may assume that φ and φ' agree on the loop γ_ℓ, and set $\varphi_C = \min\{\varphi, \varphi'\}$.   □

In three steps, we now construct a tropical independence among 20 pairwise sums of functions in

$$\mathcal{S} := \{\varphi_i : i \neq h, h+1\} \cup \{\varphi_A, \varphi_B, \varphi_C\}.$$

4.3.1 Step 1 First, we identify a collection of simpler functions in $R(D)$ that are not necessarily in Σ. Unlike φ_A and φ_B, these functions are completely explicit; they have fixed slopes at every point of the graph, rather than slopes in a restricted range. Moreover, these functions generate a tropical submodule containing φ_A, φ_B and φ_C.

Proposition 4.25 *There are functions φ_h^0, φ_{h+1}^0 and φ_h^∞ in $R(D)$ such that:*

  (i) $s_k(\varphi_h^0) = s_k[h]$ and $s'_k(\varphi_h^0) = s'_k[h]$ for all k.

 (ii) $s_k(\varphi_{h+1}^0) = s_k[h+1]$ and $s'_k(\varphi_{h+1}^0) = s'_k[h+1]$ for all k.

(iii) $s_k(\varphi_h^\infty) = s_k[h]$ and $s'_{k-1}(\varphi_h^\infty) = s'_{k-1}[h]$ for all $k \leq \ell$, and $s_k(\varphi_h^\infty) = s_k[h+1]$ and $s'_{k-1}(\varphi_h^\infty) = s'_{k-1}[h+1]$ for all $k > \ell$.

 (iv) *The function φ_A is a tropical linear combination of the functions φ_h^0 and φ_h^∞, where the two functions simultaneously achieve the minimum at a point to the right of γ_ℓ.*

  (v) *The function φ_B is a tropical linear combination of the functions φ_{h+1}^0 and φ_h^∞, where the two functions simultaneously achieve the minimum at a point to the left of γ_ℓ.*

 (vi) *The function φ_C is a tropical linear combination of the functions φ_h^0 and φ_{h+1}^0, where the two functions simultaneously achieve the minimum on the loop γ_ℓ where they agree.*

Proof The construction of the functions is essentially the same as in [19, Lemmas 11.7 and 11.19], but we describe the essential argument here, for the reader's convenience. To construct φ_h^∞, consider a function that agrees with φ_A to the left of γ_ℓ and with φ_B to the right. Because the two functions agree on γ_ℓ, they "glue" together to give a function in $R(D)$. The construction of the other two functions is similar. The verification that φ_A, φ_B and φ_C are tropical linear combinations as claimed is the same as in [19, Lemmas 11.8 and 11.19].   □

4.3.2 Step 2 Next, we choose a set $\mathcal{B}''$ of 20 pairwise sums of functions in

$$\mathcal{A} := \{\varphi_i : i \neq h, h+1\} \cup \{\varphi_h^0, \varphi_{h+1}^0, \varphi_h^\infty\}$$

that satisfies conditions (i)–(vi) of Section 4.2. We will choose this set so that, moreover, the independence θ produced by the algorithm from Section 4.2 satisfies a technical condition involving the best approximations of θ by certain functions in $R(D)$ that are not in the set (Lemma 4.30).

Start with the set $\mathcal{B}$ of pairwise sums of elements in $\mathcal{A} \setminus \{\varphi_h^\infty\}$. Note that $|\mathcal{B}| = 21$. As a first step toward specifying $\mathcal{B}''$, we choose one function $\psi \in \mathcal{B}$, of the form $\varphi_i + \varphi_j$ for $i, j \neq h, h+1$, to exclude. If $\ell \leq 7$ and $\ell \neq 6$, let ψ be such a function that is permissible on the second block. If $\ell = 6$, let $\psi \in \mathcal{B}$ be a function that is permissible on the first block. Otherwise, if $\ell > 7$, let $\psi \in \mathcal{B}$ be a function that is permissible on the third block. This choice of ψ guarantees that the number of functions in $\mathcal{B}' := \mathcal{B} \setminus \{\psi\}$ that are permissible on each block is one more than the number of loops in that block. In order to ensure a certain technical condition in the next step (Lemma 4.30), in the cases where there is some j such that $s_\ell'[h+1] + s_\ell'[j] = s_\ell(\theta) + 1$, we adjust $\mathcal{B}'$ by removing one more function and replacing it with $\varphi_h^\infty + \varphi$ for some $\varphi \in \mathcal{A}$.

Suppose there is some $\varphi \in \mathcal{A} \setminus \{\varphi_h^\infty\}$ such that $s_\ell'[h+1] + s_\ell'(\varphi) = s_\ell(\theta) + 1$. Then we define

$$\mathcal{B}'' := \mathcal{B} \cup \{\varphi_h^\infty + \varphi\} \setminus \{\varphi_h^0 + \varphi\}.$$

Otherwise, if there is no such φ, let $\mathcal{B}'' := \mathcal{B}'$.

Lemma 4.26 *The set $\mathcal{B}''$ satisfies conditions (i)–(vi) of Section 4.2, and therefore the algorithm in Section 4.2 produces an independence θ among the functions in $\mathcal{B}''$ with slopes $s_\ell(\theta)$ as specified above.*

Proof We first prove (i). First, note that, for any function $\varphi \in \mathcal{A}$, the functions $\varphi + \varphi_h^0, \varphi + \varphi_{h+1}^0$ have identical restrictions to the switching loop γ_ℓ. Because these two functions have different slopes along β_ℓ and $\beta_{\ell+1}$, however, we see that they cannot both be permissible on γ_ℓ. In the case where $\varphi_h^\infty + \varphi \in \mathcal{B}''$, we see that the restriction of this function to a loop γ_k with $k \geq \ell$ agrees with that of the function $\varphi_{h+1}^0 + \varphi$. We note, however, that since $s_\ell'[h+1] + s_\ell'(\varphi) = s_\ell(\theta) + 1$, the function $\varphi_{h+1}^0 + \varphi$ is not permissible on the loop γ_k if $k \geq \ell$.

If $\mathcal{B}'' = \mathcal{B}'$, then condition (ii) holds by the same argument as Lemma 4.10. Otherwise, note that the function in $\mathcal{B}'' \setminus \mathcal{B}'$ is permissible on the same block as the function in $\mathcal{B}' \setminus \mathcal{B}''$, so condition (ii) still holds. Condition (iii) holds because the slopes functions in $\mathcal{A}$ do not decrease from one bridge to the next. Conditions (iv)–(vi) hold vacuously. By Proposition 4.20, therefore, there is an independence ϑ among the functions in $\mathcal{B}''$. $\square$

4.3.3 Step 3 Finally, we choose a set $\mathcal{T}$ of 20 pairwise sums of functions in $\mathcal{S}$ and show that the best approximation of the θ by $\mathcal{T}$, defined as follows, is an independence.

Definition 4.27 Let $\mathcal{T}$ be a finite subset of $\mathrm{PL}(\Gamma)$. The *best approximation* of $\theta \in \mathrm{PL}(\Gamma)$ by $\mathcal{T}$ is

(24)
$$\vartheta_{\mathcal{T}} := \min\{\varphi - c(\varphi, \theta) : \varphi \in \mathcal{T}\},$$

where $c(\varphi, \theta) = \min\{\varphi(v) - \theta(v) : v \in \Gamma\}$.

Note that $\vartheta_{\mathcal{T}} \geq \theta$, and every function $\varphi \in \mathcal{T}$ achieves the minimum at some point.

Lemma 4.28 *Let $\theta = \min_{\psi \in \mathcal{B}''}\{\psi + a_\psi\}$. Suppose $\varphi = \min_{\psi' \in \mathcal{C}}\{\psi' + b_{\psi'}\}$, where $\mathcal{C} \subset \mathcal{B}''$. Then the best approximation of θ by φ achieves equality on the entire region where some $\psi' \in \mathcal{C}$ achieves the minimum in θ.*

Proof Let $c = \min_{\psi' \in \mathcal{C}}\{b_{\psi'} - a_{\psi'}\}$. Choose $\psi' \in \mathcal{C}$ such that $c = b_{\psi'} - a_{\psi'}$. Then $\varphi - c \geq \theta$, with equality at points where ψ' achieves the minimum in θ. □

We now study the best approximation of θ by various pairwise sums of function in $\mathcal{S}$.

Lemma 4.29 *Let $\varphi \in \mathcal{A} \setminus \{\varphi_h^\infty\}$. The best approximation of θ by $\varphi_C + \varphi$ achieves equality on the region where either $\varphi_h^0 + \varphi$ or $\varphi_{h+1}^0 + \varphi$ achieves the minimum.*

Proof If $\mathcal{B}''$ contains both $\varphi_h^0 + \varphi$ and $\varphi_{h+1}^0 + \varphi$, then since $\varphi_C + \varphi$ is a tropical linear combination of these two functions, the result follows from Lemma 4.28. If not, then by construction $\mathcal{B}''$ does not contain $\varphi_h^0 + \varphi$, and $s_\ell'[h+1] + s_\ell'(\varphi) = s_\ell(\theta) + 1$. In this case, $\varphi_C + \varphi$ has slope greater than $s_\ell(\theta)$ on β_ℓ, so it must achieve equality to the left of γ_ℓ, where it agrees with $\varphi_{h+1}^0 + \varphi$. □

Lemma 4.30 *Let $\varphi \in \mathcal{A} \setminus \{\varphi_h^\infty\}$. If $\varphi_h^\infty + \varphi \notin \mathcal{B}''$, then the best approximation of θ by $\varphi_h^\infty + \varphi$ achieves equality on the region where either $\varphi_h^0 + \varphi$ or $\varphi_{h+1}^0 + \varphi$ achieves the minimum.*

Proof If $\varphi_h^0 + \varphi$ is assigned to a loop γ_k with $k < \ell$, then since $\varphi_h^\infty \geq \varphi_h^0$ with equality to the left of γ_ℓ, we see that the best approximation of θ by $\varphi_h^\infty + \varphi$ achieves equality on the region where $\varphi_h^0 + \varphi$ achieves the minimum. Similarly, if $\varphi_{h+1}^0 + \varphi$ is assigned to a loop γ_k with $k \geq \ell$, then the best approximation of θ by $\varphi_h^\infty + \varphi$ achieves equality on the region where $\varphi_{h+1}^0 + \varphi$ achieves the minimum. It therefore suffices to consider the case where $\varphi_h^0 + \varphi$ is not assigned to a loop γ_k with $k < \ell$, but $\varphi_{h+1}^0 + \varphi$ is. By Lemma 4.21, on every loop γ_k in the same block as γ_ℓ with $k < \ell$, there is a departing function. It follows that

$$s_\ell[h+1] + s_\ell(\varphi) \geq s_\ell(\theta) + 1.$$

Since $\varphi_h^0 + \varphi$ is not assigned to a loop γ_k with $k < \ell$, we must have equality in the expression above. By construction, in this case $\varphi_h^\infty + \varphi \in \mathcal{B}''$. □

Remark 4.31 It is possible that the best approximation of θ by $\varphi_C + \varphi$ achieves equality on *both* the region where $\varphi_h^0 + \varphi$ achieves the minimum and the region where $\varphi_{h+1}^0 + \varphi$ achieves the minimum. However, the set of independences is open in the set of all tropical linear combinations. In other words, if the coefficients are varied in a sufficiently small neighborhood, the result is still an independence. One can therefore choose the independence θ to rule out this possibility.

We now describe our choice of the set $\mathcal{T}$. We will define sets $\mathcal{T}_j$ and $\mathcal{T}'$ below, and define

$$\mathcal{T} = \{\varphi_{ij} \in \mathcal{B}'' : i, j \neq h, h+1\} \cup \left(\bigcup_{j \neq h, h+1} \mathcal{T}_j \right) \cup \mathcal{T}'.$$

For $j \neq h, h+1$, if the best approximation of θ by $\varphi_C + \varphi_j$ achieves equality where $\varphi_h^0 + \varphi_j$ achieves the minimum, let $\mathcal{T}_j = \{\varphi_B + \varphi_j, \varphi_C + \varphi_j\}$. Otherwise, if the best approximation of θ by $\varphi_C + \varphi_j$ achieves equality where $\varphi_{h+1}^0 + \varphi_j$ achieves the minimum, then let $\mathcal{T}_j = \{\varphi_A + \varphi_j, \varphi_C + \varphi_j\}$.

Similarly, we define $\mathcal{T}'$ to be a set of three pairwise sums of elements of $\{\varphi_A, \varphi_B, \varphi_C\}$, with our choice depending on where certain functions achieve equality in the best approximation. In all cases, $\varphi_C + \varphi_C \in \mathcal{T}'$. The other functions in $\mathcal{T}'$ are determined by the following rules:

- If the best approximation of θ by $\varphi_C + \varphi_C$ achieves equality at a point to the left of γ_ℓ, then $\varphi_A + \varphi_C \in \mathcal{T}'$. Otherwise, $\varphi_B + \varphi_C \in \mathcal{T}'$.

- Suppose $\varphi_A + \varphi_C \in \mathcal{T}'$. If the best approximation of θ by $\varphi_A + \varphi_C$ achieves equality at a point to the left of γ_ℓ, then $\varphi_A + \varphi_A \in \mathcal{T}'$. Otherwise, $\varphi_A + \varphi_B \in \mathcal{T}'$.

- Suppose $\varphi_B + \varphi_C \in \mathcal{T}'$. If the best approximation of θ by $\varphi_B + \varphi_C$ achieves equality at a point to the left of γ_ℓ, then $\varphi_A + \varphi_B \in \mathcal{T}'$. Otherwise, $\varphi_B + \varphi_B \in \mathcal{T}'$.

Theorem 4.32 *The best approximation $\vartheta_{\mathcal{T}}$ is an independence, and $\vartheta_{\mathcal{T}} = \theta$ as functions.*

Proof We show that there is a bijection $F : \mathcal{T} \to \mathcal{B}''$ with the property that each $\psi \in \mathcal{T}$ achieves the minimum in $\vartheta_{\mathcal{T}}$ on exactly the same region where $F(\psi)$ achieves the minimum in θ. From this it follows that $\vartheta_{\mathcal{T}}$ is an independence, and that $\vartheta_{\mathcal{T}} = \theta$.

For $i, j \neq h, h+1$, we set $F(\varphi_{ij}) = \varphi_{ij}$. Next, consider a value $j \neq h, h+1$. We describe the restriction of F to the subset $\mathcal{T}_j$. The restriction of F to $\mathcal{T}'$ admits a similar description. By Lemma 4.29, the best approximation of θ by $\varphi_C + \varphi_j$ achieves equality on the region where either $\varphi_h^0 + \varphi_j$ or $\varphi_{h+1}^0 + \varphi_j$ achieves the minimum (but not both, see Remark 4.31). If it achieves equality on the region where $\varphi_h^0 + \varphi_j$ achieves the minimum, set $F(\varphi_C + \varphi_j) = \varphi_h^0 + \varphi_j$. Otherwise, set $F(\varphi_C + \varphi_j) = \varphi_{h+1}^0 + \varphi_j$.

Suppose that $F(\varphi_C + \varphi_j) = \varphi_{h+1}^0 + \varphi_j$. The case where $F(\varphi_C + \varphi_j) = \varphi_h^0 + \varphi_j$ follows from a similar (in fact, simpler) argument. Since φ_C agrees with φ_{h+1}^0 at points on or to the left of γ_ℓ, we have $\varphi_A + \varphi_j \in \mathcal{T}$. If $\varphi_h^\infty + \varphi_j \in \mathcal{B}''$, then we set $F(\varphi_A + \varphi_j) = \varphi_h^\infty + \varphi_j$. In this case, we have $s_\ell'[h] + s_\ell'[j] = s_\ell(\theta)$. Since

γ_ℓ is the last loop in its block, we see that the slope of $\varphi_A + \varphi_j$ is greater than that of θ on the right half of $\beta_{\ell+1}$. Thus, the best approximation of θ by $\varphi_A + \varphi_j$ must achieve equality to the left of $\beta_{\ell+1}$, where $\varphi_A + \varphi_j$ agrees with $\varphi_h^\infty + \varphi_j$.

If $\varphi_h^\infty + \varphi_j \notin \mathcal{B}''$, then set $F(\varphi_A + \varphi_j) = \varphi_h^0 + \varphi_j$, and consider the best approximation θ' of θ by $\mathcal{B}'' \cup \{\varphi_h^\infty + \varphi_j\}$. Note that the coefficient of $\varphi_A + \varphi_j$ is the same in the best approximation of θ' by $\varphi_A + \varphi_j$ and the best approximation of θ by $\varphi_A + \varphi_j$. By Lemma 4.30, $\varphi_h^\infty + \varphi_j$ achieves equality in θ' on the region where either $\varphi_h^0 + \varphi_j$ or $\varphi_{h+1}^0 + \varphi_j$ achieves the minimum in θ. Then, since φ_A is a linear combination of φ_h^0 and φ_h^∞, by Lemma 4.28, it follows that the best approximation of θ by $\varphi_A + \varphi_j$ achieves equality on the region where either $\varphi_h^0 + \varphi_j$ or $\varphi_{h+1}^0 + \varphi_j$ achieves the minimum. But φ_A and φ_C do not agree at any point to the left of γ_ℓ, so the best approximation of θ by $\varphi_A + \varphi_j$ must achieve equality on the region where either $\varphi_h^0 + \varphi_j$ achieves the minimum. $\qquad\square$

5 Effectivity of the virtual class

Recall that $\widetilde{\mathfrak{M}}_{13}$ is an open substack of the moduli stack of stable curves, and $\widetilde{\mathfrak{S}}_d^r$ is a stack of generalized limit linear series of rank r and degree d over $\widetilde{\mathfrak{M}}_{13}$. There is a morphism of vector bundles $\phi \colon \operatorname{Sym}^2(\mathcal{E}) \to \mathcal{F}$ over $\widetilde{\mathfrak{S}}_d^r$, whose degeneracy locus is denoted by $\mathfrak{U}$.

The case of Theorem 4.2 where $g = 13$ shows that the pushforward $\sigma_*[\mathfrak{U}]^{\mathrm{virt}}$ under the proper forgetful map $\sigma \colon \widetilde{\mathfrak{S}}_d^r \to \widetilde{\mathcal{M}}_g$ is a divisor, not just a divisor class. In our proof that $\sigma_*[\mathfrak{U}]^{\mathrm{virt}}$ is effective, we will use the additional cases where $g = 11$ or 12. Theorem 4.2 implies the following result.

Theorem 5.1 *Let X be a general curve of genus $g \in \{11, 12, 13\}$, and let $p \in X$ be a general point. Let $V \subseteq H^0(X, L)$ be a linear series of degree 16 and rank 5. Assume that*

(i) *if $g = 12$, then $a_1^V(p) \geq 2$, and*

(ii) *if $g = 11$, then either $a_1^V(p) \geq 3$, or $a_0^V(p) + a_2^V(p) \geq 5$.*

Then the multiplication map $\phi_V \colon \operatorname{Sym}^2 V \to H^0(X, L^{\otimes 2})$ is surjective.

We now prove that $\mathfrak{U}$ is generically finite over each component of $\sigma_*[\mathfrak{U}]^{\mathrm{virt}}$, which implies that $\sigma_*[\mathfrak{U}]^{\mathrm{virt}}$ is effective. Our argument follows closely that of [19, Section 12]. Indeed, several of the lemmas and propositions along the way are identical, and we omit those proofs. As in [19, Section 12], we suppose that $Z \subseteq \overline{\mathcal{M}}_{13}$ is an irreducible divisor and that $\sigma|_{\mathfrak{U}}$ has positive-dimensional fibers over the generic point of Z. Let $j_2 \colon \overline{\mathcal{M}}_{2,1} \to \overline{\mathcal{M}}_{13}$ be the map obtained by attaching an arbitrary pointed curve of genus 2 to a fixed general pointed curve (X, p) of genus 11. Since $g = 13$ is odd, by [19, Proposition 2.2], it suffices to show the following:

(a) Z is the closure of a divisor in $\mathcal{M}_{13}$,

(b) $j_2^*(Z) = 0$, and

(c) Z does not contain any codimension 2 stratum $\Delta_{2,j}$.

The only irreducible boundary divisors in $\widetilde{\mathcal{M}}_{13}$ are Δ_0° and Δ_1°. Therefore, item (a), that Z is the closure of a divisor in $\mathcal{M}_{13}$, is a consequence of the following.

Proposition 5.2 *The image of the degeneracy locus $\mathfrak{U}$ does not contain Δ_0° or Δ_1°.*

Proof The proof is identical to [19, Proposition 12.3]. □

The proofs of (b) and (c) use the following lemma.

Lemma 5.3 *If $[X] \in Z$ and $p \in X$, then there is a linear series $V \in G_{16}^5(X)$ that is ramified at p such that ϕ_V is not surjective.*

Proof The proof is identical to [19, Lemma 12.4]. □

5.1 Pulling back to $\overline{\mathcal{M}}_{2,1}$

In order to verify (b), we consider the preimage of Z under the map J_2.

Lemma 5.4 *The preimage $J_2^{-1}(Z)$ is contained in the Weierstrass divisor $\overline{\mathcal{W}}_2$ in $\overline{\mathcal{M}}_{2,1}$.*

Proof The proof is identical to [19, Lemma 12.5]. □

To prove that $J_2^*(Z) = 0$, we consider the following construction. Let Γ be a chain of 13 loops with the following restrictions on edge lengths:

  (i) $m_2 = \ell_2$ (that is, the second loop has torsion index 2),

  (ii) $n_3 \gg n_2$, and

  (iii) $\ell_{k+1} \ll m_k \ll \ell_k \ll n_{k+1} \ll n_k$ for all $k \neq 2$.

The last condition says that, subject to the constraints of conditions (i) and (ii), the edge lengths otherwise satisfy (20). Let X be a smooth curve of genus 13 over K whose skeleton is Γ. We first note the following.

Lemma 5.5 *If $[X] \notin Z$, then $J_2^*(Z) = 0$.*

Proof This proof is identical to the first part of the proof of [19, Proposition 12.6]. □

Proposition 5.6 *We have $J_2^*(Z) = 0$.*

Proof By Lemma 5.5, it suffices to show that $[X] \notin Z$. We divide Γ into two subgraphs $\widetilde{\Gamma}'$ and $\widetilde{\Gamma}$, to the left and right, respectively, of the midpoint of the long bridge β_3. Let $q \in X$ be a point specializing to v_{14}. If $[X] \in Z$, by Lemma 5.3 there is a linear series in the degeneracy locus over X that is ramified at q. We now show that this impossible.

Let $\ell = (L, V) \in G^5_{16}(X)$ be a linear series ramified at q. We may assume that $L = \mathcal{O}(D_X)$, where $D = \mathrm{Trop}(D_X)$ is a break divisor, and consider $\Sigma = \mathrm{trop}(V)$. We will show that there are 20 tropically independent pairwise sums of functions in Σ using a variant of the arguments in Section 4. It follows that the multiplication map ϕ_ℓ is surjective, and hence $[X]$ cannot be in Z.

To produce 20 tropically independent pairwise sums of functions in Σ, following the methods of Section 4, we first consider the slope sequence along the long bridge β_3. First, suppose that either $s_3[4] \leq 2$ or $s_3[3] + s_3[5] \leq 5$. In this case, even though the restriction of Σ to $\widetilde{\Gamma}$ is not the tropicalization of a linear series on a pointed curve of genus 11 with prescribed ramification, it satisfies all of the combinatorial properties of the tropicalization of such a linear series. The proof of Theorem 4.2 then goes through verbatim, yielding a tropical linear combination of 20 functions in Σ such that each function achieves the minimum uniquely at some point of $\widetilde{\Gamma} \subseteq \Gamma$.

For the remainder of the proof, we therefore assume that $s_3[4] \geq 3$ and $s_3[3] + s_3[5] \geq 6$. Since $\deg D_{|\widetilde{\Gamma}'} = 5$, we see that $s_3[5] \leq 5$. Moreover, since the divisor $D_{|\widetilde{\Gamma}'} - s_3[4]w_2$ has positive rank on $\widetilde{\Gamma}'$, and no divisor of degree one on $\widetilde{\Gamma}'$ has positive rank, $s_3[4]$ must be exactly 3. Since the canonical class is the only divisor class of degree two and rank one on $\widetilde{\Gamma}'$, we see that $D_{|\widetilde{\Gamma}'} \sim K_{\widetilde{\Gamma}'} + 3w_2$. This yields an upper bound on each of the slopes $s_3[i]$, and these bounds determine the slopes for $i \geq 2$:

$$s_3[5] = 5, \quad s_3[4] = 3, \quad s_3[3] = 1, \quad s_3[2] = 0.$$

Moreover, we must have $s_2'[i] = s_3[i]$ for $2 \leq i \leq 5$. Since ℓ is ramified at q, we also have $s_{14}[5] \geq 6$. These conditions together imply that the sum of the multiplicities of all loops and bridges on $\widetilde{\Gamma}$ is at most one.

To construct an independence on Γ, we first construct an independence among 5 functions on $\widetilde{\Gamma}'$. This is done exactly as in [19, Figure 39], and we omit the details.

Next, we construct an independence among 15 pairwise sums of functions in Σ restricted to $\widetilde{\Gamma}$, with the property that any function ψ that obtains the minimum on $\widetilde{\Gamma}$ satisfies $s_2'(\psi) \leq 4$. Note that each of the functions ψ that obtains the minimum on $\widetilde{\Gamma}'$ satisfies $s_3(\psi) \geq 5$. Since the bridge β_3 is very long, it follows that no function that obtains the minimum on one of the two subgraphs can obtain the minimum on the other. Thus, we have constructed a tropical linear combination of 20 pairwise sums of functions in Σ in which 5 achieve the minimum uniquely at some point of $\widetilde{\Gamma}'$ and 15 achieve the minimum uniquely at some point of $\widetilde{\Gamma}$. In particular, this is an independence, as required.

It remains to construct an independence among 15 pairwise sums of functions in Σ restricted to $\widetilde{\Gamma}$. To do this, we run the algorithm from [19], with one change. (Indeed, one could imagine that Γ is simply the first 13 loops in a chain of 23 loops; we construct the independence from [19, Section 12.3], and restrict it to Γ.) At the start, we skip the step named "start at the first bridge". Instead, we do not assign any function ψ with $s_3(\psi) \geq 5$, and we start with the loop subroutine applied to γ_3. Following this construction, there will only be two blocks, and there will be two functions with slope 2 along the last

bridge β_{14}. We eliminate one of these functions from $\mathcal{B}$, and assign the other to β_{14}. The rest of the argument is exactly the same as that of [19]. □

5.2 Higher-codimension boundary strata

It remains to verify (c), that Z does not contain any of the codimension 2 boundary strata $\Delta_{2,j} \subseteq \overline{\mathcal{M}}_{13}$.

Proposition 5.7 *The component Z does not contain any codimension 2 stratum $\Delta_{2,j}$.*

Proof The proof is again a variation on the independence constructions from the proof of Theorem 4.2. We fix $\ell = 11 - j$. Let Y_1 be a smooth curve of genus 2 over K whose skeleton Γ_1 is a chain of 2 loops with bridges, and let $p \in Y_1$ be a point specializing to the right endpoint of Γ_1. Similarly, let Y_2 and Y_3 be smooth curves of genus ℓ and j, respectively, whose skeletons Γ_2 and Γ_3, are chains of ℓ loops and j loops with edge lengths satisfying (20). Suppose further that the edges in the final loop of Γ_2 are much longer than those in the first loop of Γ_3. Let $p, q \in Y_2$ be points specializing to the left and right endpoints of Γ_2, respectively, and let $q \in Y_3$ be a point specializing to the left endpoint of Γ_3. We show that $[Y'] = [Y_1 \cup_p Y_2 \cup_q Y_3] \in \Delta_{2,j}$ is not contained in Z.

As in the proof of [19, Proposition 12.6], if $[Y'] \in Z$, then Z contains points $[X]$ corresponding to smooth curves whose skeletons are arbitrarily close to the skeleton of Y' in the natural topology on $\overline{M}_{13}^{\text{trop}}$. In particular, there is an $X \in Z$ with skeleton a chain of loops Γ_X whose edge lengths satisfy all the conditions of (20), except that the bridges β_3 and β_ℓ are exceedingly long in comparison to the other edges. Let Γ be the subgraph of Γ_X to the right of the midpoint of the bridge β_3. Note that Γ is a chain of 11 loops, labeled $\gamma_3, \ldots, \gamma_{13}$, with bridges labeled $\beta_3, \ldots, \beta_{14}$.

By Lemma 5.3, there is a linear series V of degree 16 and rank 5 on X that is ramified at a point x specializing to the right-hand endpoint v_{14}, and such that ϕ_V is not surjective. We will show that this is not possible, using the tropical independence construction from Section 4. Let $\Sigma = \text{trop}(V)$. We have that either $s_2'[4] \leq 2$ or $s_2'[3] + s_2'[4] \leq 5$. Also, since V is ramified at x, we have $s_{14}[5] \geq 6$. These conditions imply that the multiplicity of every loop and bridge is zero. In particular, for each i there is a function φ_i satisfying

$$s_k(\varphi_i) = s_{k-1}'(\varphi_i) = s_k[i] = s_{k-1}'[i] \quad \text{for all } k.$$

These functions have constant slope along bridges, and the slopes $s_k(\varphi_i)$ are nondecreasing in k. These properties guarantee that, even though the bridge β_ℓ is very long, a function φ_{ij} can only obtain the minimum on a loop or bridge where it is permissible.

Even though the restriction of Σ to Γ is not the tropicalization of a linear series on a curve of genus 11 with prescribed ramification at two specified points specializing to the left and right endpoints of Γ, it satisfies all of the combinatorial properties of the tropicalization of such a linear series, and we may apply the algorithm from Section 4. Because we are in a situation where the relative lengths of the bridges do not matter (Remark 4.17) the construction yields an independence among 20 pairwise sums of functions in Σ, and the proposition follows. □

6 The Bertram–Feinberg–Mukai conjecture in genus 13

The aim of this section is to prove the existence part of the Bertram–Feinberg–Mukai conjecture on $\overline{\mathcal{M}}_{13}$. For a smooth curve X of genus g, we denote by $\mathrm{SU}_X(2, \omega)$ the moduli space of S–equivalence classes of semistable rank-two vector bundles E on X with $\det(E) \cong \omega_X$. For an integer $k \geq 0$, the Brill–Noether locus

$$\mathrm{SU}_X(2, \omega, k) := \{E \in \mathrm{SU}_X(2, \omega) : h^0(X, E) \geq k\}$$

has the structure of a Lagrangian degeneracy locus and each component of $\mathrm{SU}_X(2, \omega, k)$ has dimension at least

$$\beta(2, g, k) = 3g - 3 - \binom{k+1}{2};$$

see [38]. Furthermore, $\mathrm{SU}_X(2, \omega, k)$ is smooth of dimension $\beta(2, g, k)$ at a point $[E]$ corresponding to a stable vector bundle if and only if the Mukai–Petri map (1) is injective. Of particular interest to us is the case

$$g = 13 \quad \text{and} \quad k = 8,$$

in which case $\beta(2, 13, 8) = 0$. First, using linkage methods, we show that a general curve of genus 13 carries a stable vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$. Then using a Hecke correspondence, we compute the fundamental class of $\mathrm{SU}_X(2, \omega, 8)$.

Theorem 6.1 *A general curve X of genus* 13 *caries a stable vector bundle E of rank two with* $\det E \cong \omega_X$ *and* $h^0(X, E) = 8$.

As a first step towards proving Theorem 6.1, we determine the extension type of the vector bundles in question.

Proposition 6.2 *For a general curve X of genus* 13*, every vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$ can be represented as an extension*

$$(25) \qquad\qquad 0 \to \mathcal{O}_X(D) \to E \to \omega_X(-D) \to 0,$$

where D is an effective divisor of degree 6 on X, such that $L := \omega_X(-D) \in W_{18}^6(X)$ is very ample and the map $\phi_L : \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$ is not surjective. Conversely, a very ample $L \in W_{18}^6(X)$ with ϕ_L not surjective induces a stable vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$.

Proof Using a result of Segre — see [39] or [33, Proposition 3.1] for modern proofs — every semistable vector bundle E on X of rank two and canonical determinant carries a line subbundle $\mathcal{O}_X(D) \hookrightarrow E$ with $\deg D \geq \frac{1}{2}(g-2)$. Therefore, in our case $\deg D \geq 6$.

If $h^0(X, \mathcal{O}_X(D)) \geq 2$, since $h^0(X, \mathcal{O}_X(D)) + h^0(X, \omega_X(-D)) \geq h^0(X, E) = 8$ it follows from the Brill–Noether theorem and Riemann–Roch that $\deg D = 8$, hence $\omega_X(-D) \in W_{16}^5(X)$. It follows that the extension (25) lies in the kernel of the map

$$\mathrm{Ext}^1(\omega_X(-D), D) \to H^0(\omega_X(-D))^\vee \otimes H^1(D).$$

This implies that the multiplication map $\phi_{\omega_X(-D)}\colon \operatorname{Sym}^2 H^0(X, \omega_X(-D)) \to H^0(X, \omega_X^{\otimes 2}(-2D))$ is not surjective, which contradicts Theorem 1.5. Therefore $h^0(X, \mathcal{O}_X(D)) = 1$, in which case necessarily $\deg D = 6$ and $h^0(X, E) = h^0(X, \mathcal{O}_X(D)) + h^0(X, \omega_X(-D))$. Setting $L := \omega_X(-D) \in W_{18}^6(X)$, an extension E satisfies $h^0(X, E) = 8$ if and only if the extension class of E in $\operatorname{Ext}^1(L, D)$ lies in the kernel of the linear map

$$\operatorname{Ext}^1(L, D) \to H^0(L)^\vee \otimes H^1(D).$$

Thus, an extension (25) exists if and only if the multiplication map

$$\phi_L \colon \operatorname{Sym}^2 H^0(L) \to H^0(X, L^{\otimes 2}) \cong \operatorname{Ext}^1(L, D)^\vee$$

is not surjective. We claim that L is very ample. Otherwise, there exist points $x, y \in X$ such that $L' := L(-x - y) \in W_{16}^5(X)$. Since X is general, by Theorem 1.5 the multiplication map

$$\phi_{L'} \colon \operatorname{Sym}^2 H^0(X, L') \to H^0(X, (L')^{\otimes 2})$$

is surjective, implying the inclusion $H^0(X, (L')^{\otimes 2}(x + y)) \subseteq \operatorname{Im}(\phi_L)$. We deduce that $[E]$ lies in the kernel of the map

$$\operatorname{Ext}^1(L, D) \to \operatorname{Ext}^1(L(-x - y), D).$$

That is, the vector bundle E can also be represented as an extension

$$0 \to L(-x - y) \to E \to \mathcal{O}_X(D + x + y) \to 0,$$

thus contradicting the semistability of E. We conclude that L has to be very ample.

Conversely, each *very ample* linear system $L \in W_{18}^6(X)$, for which the map ϕ_L is not surjective induces a stable vector bundle E; see also [14, 7.2]. Indeed, let us assume E is not semistable. In view of the extension (25), a maximally destabilizing line subbundle of E is of the form $L(-M)$, where M is an effective divisor on X with $\deg M \leq 6$. Therefore, apart from (25), E can also be realized as an extension

$$0 \to L(-M) \to E \to \mathcal{O}_X(D + M) \to 0.$$

By applying Riemann–Roch, one can then write

$$h^0(X, L(-M)) + h^1(X, L(-M)) = h^0(X, L) + h^1(X, L) - 2 \dim \frac{H^0(X, L)}{H^0(X, L(-M))} + \deg(M).$$

Since

$$h^0(X, L) + h^1(X, L) = h^0(X, E) \leq h^0(X, L(-M)) + h^1(X, L(-M)),$$

it follows that

$$\deg M \geq 2 \dim \frac{H^0(L)}{H^0(L(-M))}.$$

Since L is very ample, we find $\deg M \in \{4, 5, 6\}$. In each case, the Brill–Noether number of $L(-M)$ is negative, contradicting the generality of X. Therefore E is stable. $\qquad\square$

Proof of Theorem 6.1 By Proposition 6.2, it suffices to show that for a general curve X of genus 13, there exists a very ample linear system $L \in W^6_{18}(X)$ such that ϕ_L is not surjective. We use a method inspired by Verra's proof [47] of the unirationality of $\overline{\mathcal{M}}_{14}$. To illustrate the idea behind the proof, first suppose that there exists an embedding $\varphi_L \colon X \hookrightarrow \mathbb{P}^6$ given by $L \in W^6_{18}(X)$, such that the map ϕ_L is not surjective. In particular, $X \subseteq \mathbb{P}^6$ lies on at least $5 = \binom{8}{2} - h^0(X, L^{\otimes 2}) - 1$ quadrics. We expect the base locus of this system of quadrics to be a reducible curve (of degree 32), containing X as a component and accordingly write

$$X + C = \mathrm{Bs}\, |\mathcal{I}_{X/\mathbb{P}^6}(2)|.$$

Assuming that X and C intersect transversally, we obtain that $X + C$ is a complete intersection curve in $\mathbb{P}^6$. Therefore C is a curve of degree $14 = 2^5 - \deg(X)$ and applying the adjunction formula $2g(X) - 2g(C) = (10 - 7)(\deg(X) - \deg(C)) = 12$ (see for instance [47, page 1429]), we obtain $g(C) = 7$.

We now reverse this procedure and start with a general curve $C \subseteq \mathbb{P}^6$ of genus 7 embedded by a 7–dimensional linear system $V \subseteq H^0(C, L_C)$, where $L_C \in \mathrm{Pic}^{14}(C)$ is a general line bundle, therefore $h^0(C, L_C) = 8$. Consider the multiplication map

$$\phi_V \colon \mathrm{Sym}^2(V) \to H^0(C, L_C^{\otimes 2})$$

and observe that $\mathrm{Ker}(\phi_V)$ has dimension at least $6 = \dim \mathrm{Sym}^2(V) - h^0(L_C^{\otimes 2})$. Choose a general 5–dimensional system of quadrics $W \in G\big(5, H^0(\mathbb{P}^6, \mathcal{I}_{C/\mathbb{P}^6}(2))\big)$. We then expect

$$(26) \qquad\qquad \mathrm{Bs}\, |W| = C + X \subseteq \mathbb{P}^6$$

to be a nodal curve, and the curve X linked to C to be a smooth curve of degree 18 and genus 13. Setting $L := \mathcal{O}_X(1) \in W^6_{18}(X)$, by construction L is very ample and the embedded curve $X \subseteq \mathbb{P}^6$ lies on at least 5 quadrics, therefore ϕ_L is not surjective.

To carry this out, one needs to check some transversality statements. Let $\mathcal{P}ic^{14}_7$ be the universal Picard variety parametrizing pairs $[C, L_C]$, where C is a smooth curve of genus 7 and $L_C \in \mathrm{Pic}^{14}(C)$. As pointed out in [47, Theorem 1.2], it follows from Mukai's work [37] that $\mathcal{P}ic^{14}_7$ is unirational. We introduce the variety

$$\mathcal{Y} := \{[C, L_C, V, W] : [C, L_C] \in \mathcal{P}ic^{14}_7,\, V \in G(6, H^0(C, L_C)),\, W \in G(5, \mathrm{Ker}(\phi_V))\}$$

The forgetful map $\mathcal{Y} \to \mathcal{P}ic^{14}_7$ has the structure of an iterated locally trivial projective bundle over $\mathcal{P}ic^{14}_7$, therefore $\mathcal{Y}$ is unirational as well. Moreover,

$$\dim \mathcal{Y} = \dim \mathcal{P}ic^{14}_7 + \dim G(7,8) + \dim G(5,6) = 4 \cdot 7 - 3 + 7 + 5 = 37.$$

One has a rational *linkage map*

$$\chi \colon \mathcal{Y} \dashrightarrow \mathcal{SU}_{13}(2, \omega, 8), \quad [C, L_C, V, W] \mapsto [X, L, E],$$

where X is defined by (26), $L := \mathcal{O}_X(1) \in W_{18}^6(X)$ and $E \in \mathrm{SU}_X(2, \omega, 8)$ is the rank-two vector bundle defined uniquely by the extension $0 \to \omega_X \otimes L^\vee \to E \to L \to 0$.

To show that χ is well defined it suffices to produce one example of a point in $\mathcal{Y}$ for which all these assumptions are realized. To that end, we consider 11 general points $p_1, \ldots, p_5$ and $q_1, \ldots, q_6$ respectively in $\mathbb{P}^2$ and the linear system

$$H \equiv 6h - 2(E_{p_1} + \cdots + E_{p_5}) - (E_{q_1} + \cdots + E_{q_6})$$

on the blowup $S = \mathrm{Bl}_{11}(\mathbb{P}^2)$ at these points. Here h denotes the pullback of the line class from $\mathbb{P}^2$. Via *Macaulay2* one checks that $S \xhookrightarrow{|H|} \mathbb{P}^6$ is an embedding and the graded Betti diagram of S is

$$
\begin{array}{ccccc}
1 & - & - & - & - \\
- & 5 & - & - & - \\
- & - & 15 & 16 & 15
\end{array}
$$

Next we consider a general curve $C \subseteq S$ in the linear system

$$C \equiv 10h - 4(E_{p_1} + E_{p_2} + E_{p_3} + E_{p_4}) - 3E_{p_5} - 2(E_{q_1} + E_{q_2}) - (E_{q_3} + E_{q_4} + E_{q_5} + E_{q_6}).$$

Via *Macaulay2*, we verify that C is smooth, $g(C) = 7$ and $\deg(C) = 14$. Furthermore, using that $H^1(\mathbb{P}^6, \mathcal{I}_{S/\mathbb{P}^6}(2)) = 0$, we have an exact sequence

$$0 \to H^0(\mathbb{P}^6, \mathcal{I}_{S/\mathbb{P}^6}(2)) \to H^0(\mathbb{P}^6, \mathcal{I}_{C/\mathbb{P}^6}(2)) \to H^0(S, \mathcal{O}_S(2H - C)) \to 0.$$

Since $\mathcal{O}_S(2H - C) = \mathcal{O}_S(2h - E_{p_5} - E_{q_3} - E_{q_4} - E_{q_5} - E_{q_6})$, clearly $h^0(S, \mathcal{O}_S(2H - C)) = 1$, therefore $h^0(\mathbb{P}^6, \mathcal{I}_{C/\mathbb{P}^6}(2)) = 6$. That is, $C \subseteq \mathbb{P}^6$ is a 2–normal curve.

One also verifies with *Macaulay2* that $C \subseteq \mathbb{P}^6$ is scheme-theoretically cut out by quadrics. Using [47, Proposition 2.2], C lies on a smooth surface $Y \subseteq \mathbb{P}^6$ which is a complete intersection of four quadrics containing C. Furthermore, the linear system $|\mathcal{O}_Y(2H - C)|$ is basepoint-free, so a general element $X \in |\mathcal{O}_Y(2H - C)|$ is a smooth curve of genus 13 meeting C transversally. Finally, a standard argument using the exact sequence $0 \to \mathcal{O}_Y(H - X) \to \mathcal{O}_Y(H) \to \mathcal{O}_X(H) \to 0$ shows that since C is 2–normal, the residual curve X is 1–normal. That is, $h^1(X, \mathcal{O}_X(1)) = 1$. This implies that the map $\chi : \mathcal{Y} \dashrightarrow \mathcal{SU}_{13}(2, \omega, 8)$ is well defined and dominant. $\qquad\square$

Corollary 6.3 *The parameter space $\mathcal{SU}_{13}(2, \omega, 8)$ is unirational.*

Proof This follows from the proof of Theorem 6.1 and from the unirationality of $\mathcal{Y}$. $\qquad\square$

6.1 The fundamental class of $\mathrm{SU}_X(2, \omega, 8)$ for a general curve

It is essential for our calculations to determine the degree of the map

$$\vartheta : \mathcal{SU}_{13}(2, \omega, 8) \to \mathcal{M}_{13}, \quad \vartheta([X, E]) = [X].$$

We fix a general curve X of genus g and a point $p \in X$. Since the moduli space $\mathrm{SU}_X(2, \omega)$ is singular, in order to determine the fundamental class of the nonabelian Brill–Noether locus $\mathrm{SU}_X(2, \omega, k)$, following [40; 33; 38] one uses instead the Hecke correspondence relating $\mathrm{SU}_X(2, \omega)$ to the smooth moduli space $\mathrm{SU}_X(2, \omega(p))$ of stable rank-two vector bundles F on X with $\det(F) \cong \omega_X(p)$.

Recall that $\mathrm{SU}_X(2, \omega(p))$ is a fine moduli space. Hence there is a universal rank-two vector bundle $\mathcal{F}$ on $X \times \mathrm{SU}_X(2, \omega(p))$ and we consider the *Hecke correspondence*

$$\boldsymbol{P} := \mathbb{P}(\mathcal{F}_{|\{p\} \times \mathrm{SU}_X(2, \omega(p))}),$$

endowed with the projection $\pi_1 \colon \boldsymbol{P} \to \mathrm{SU}_X(2, \omega(p))$. The points of $\boldsymbol{P}$ are exact sequences

$$(27) \qquad\qquad\qquad\qquad 0 \to E \to F \to K(p) \to 0,$$

where $F \in \mathrm{SU}_X(2, \omega(p))$, and therefore $\det(E) \cong \omega_X$. One has a diagram

$$
\begin{array}{ccc}
 & \boldsymbol{P} & \\
{\scriptstyle \pi_1} \swarrow & & \searrow {\scriptstyle \rho} \\
\mathrm{SU}_X(2, \omega(p)) & & \mathrm{SU}_X(2, \omega)
\end{array}
$$

where ρ assigns to a sequence (27) the semistable vector bundle E. Set

$$h := c_1(\mathcal{O}_{\boldsymbol{P}}(1)) = \rho^* c_1(\mathcal{L}_{\mathrm{ev}}),$$

where $\mathcal{L}_{\mathrm{ev}}$ is the determinant line bundle on $\mathrm{SU}_X(2, \omega)$, associated to the effective divisor

$$\Theta := \{E \in \mathrm{SU}_X(2, \omega) : H^0(X, E) \neq 0\}.$$

Set $\alpha := c_1(\mathcal{L}_{\mathrm{odd}}) \in H^2\big(\mathrm{SU}_X(2, \omega(p)), \mathbb{Z}\big)$, where $\mathcal{L}_{\mathrm{odd}}$ is the ample generator of $\mathrm{Pic}\big(\mathrm{SU}_X(2, \omega(p))\big)$. Note that $\mathrm{Pic}(\boldsymbol{P})$ is generated by h and by $\pi_1^*(\alpha)$.

For each $k \in \mathbb{N}$, the nonabelian Brill–Noether locus

$$B_{\boldsymbol{P}}(k) := \{[0 \to E \to F \to K(p) \to 0] \in \boldsymbol{P} : h^0(X, E) \geq k\}$$

has the structure of a Lagrangian degeneracy locus of expected codimension $\beta(2, g, k) + 1 = 3g - 2 - \binom{k+1}{2}$; see [38, Section 5; 33, Section 2]. As such, its virtual class $[B_{\boldsymbol{P}}(k)]^{\mathrm{virt}} \in H^*(\boldsymbol{P}, \mathbb{Q})$ can be computed in terms of certain tautological classes, whose definition we recall now.

Following [40], we consider the Künneth decomposition of the Chern classes of $\mathcal{F}$, using that $\det(\mathcal{F}) \cong \omega_X(p) \boxtimes \mathcal{L}_{\mathrm{odd}}$, and write

$$c_1(\mathcal{F}) = \alpha + (2g - 1)\varphi \quad \text{and} \quad c_2(\mathcal{F}) = \chi + \psi + g\alpha \otimes \varphi,$$

where $\varphi \in H^2(X, \mathbb{Q})$ is the fundamental class of the curve, $\chi \in H^4\big(\mathrm{SU}_X(2, \omega(p)), \mathbb{Q}\big)$ and ψ is in $H^3\big(\mathrm{SU}_X(2, \omega(p)), \mathbb{Q}\big) \otimes H^1(X, \mathbb{Q})$. Finally, we define the class

$$\gamma \in H^6\big(\mathrm{SU}_X(2, \omega(p)), \mathbb{Q}\big)$$

by the formula $\psi^2 = \gamma \otimes \varphi$. One has the relation

$$h^2 = \alpha h - \tfrac{1}{4}(\alpha^2 - \beta) \in H^4(\boldsymbol{P}, \mathbb{Q}),$$

from which we can recursively determine all powers of h. We summarize as follows.

Proposition 6.4 *For each $n \geq 2$, the following relation holds in $H^*(\boldsymbol{P}, \mathbb{Q})$:*

$$h^n = \frac{h(-2\alpha + 2h)\sqrt{\beta} + \alpha^2 - 2\alpha h + \beta}{\sqrt{\beta}(\alpha^2 - \beta)}\left(\frac{\alpha + \sqrt{\beta}}{2}\right)^n + \frac{h(2\alpha - 2h)\sqrt{\beta} + \alpha^2 - 2\alpha h + \beta}{\sqrt{\beta}(\alpha^2 - \beta)}\left(\frac{\alpha - \sqrt{\beta}}{2}\right)^n.$$

In this formula $\sqrt{\beta}$ is a formal root of the class β. Applying [33, Section 3] or [38], one can endow $B_{\boldsymbol{P}}(k)$ with the structure of a Lagrangian degeneracy locus as follows. Let $\mathcal{E}$ be the vector bundle on $X \times \boldsymbol{P}$ defined by the exact sequence

$$0 \to \mathcal{E} \to (\mathrm{id} \times \pi_1)^*(\mathcal{F}) \to (p_2)_*(\mathcal{O}_{\boldsymbol{P}}(1)) \to 0,$$

where $p_2 \colon X \times \boldsymbol{P} \to \boldsymbol{P}$ is the projection. Choose an effective divisor D of large degree on X and also denote by D its pullback under $X \times \boldsymbol{P} \to X$. Then $(p_2)_*(\mathcal{E}/\mathcal{E}(-D))$ and $(p_2)_*(\mathcal{E}(D))$ are Lagrangian subbundles of $(p_2)_*(\mathcal{E}(D)/\mathcal{E}(-D))$. For each point $t := [0 \to E \to F \to K(p) \to 0] \in \boldsymbol{P}$, one has

$$(p_2)_*(\mathcal{E}(D))(t) \cap (p_2)_*(\mathcal{E}/\mathcal{E}(-D))(t) \cong H^0(X, E).$$

Assume from now on $g = 13$ and $k = 8$, therefore we expect $B_{\boldsymbol{P}}(8)$ to be one-dimensional. Applying the formalism for Lagrangian degeneracy loci [38, Proposition 1.11], we find the following determinantal formula for its virtual fundamental class:

(28)
$$[B_{\boldsymbol{P}}(8)]^{\mathrm{virt}} = \begin{vmatrix} c_8 & c_9 & c_{10} & c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \\ c_6 & c_7 & c_8 & c_9 & c_{10} & c_{11} & c_{12} & c_{13} \\ c_4 & c_5 & c_6 & c_7 & c_8 & c_9 & c_{10} & c_{11} \\ c_2 & c_3 & c_4 & c_5 & c_6 & c_7 & c_8 & c_9 \\ c_0 & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7 \\ 0 & 0 & c_0 & c_1 & c_2 & c_3 & c_4 & c_5 \\ 0 & 0 & 0 & 0 & c_0 & c_1 & c_2 & c_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & c_0 & c_1 \end{vmatrix},$$

where the $c_i \in H^{2i}(\boldsymbol{P}, \mathbb{Q})$ are defined recursively by the following formulas, see [33, Corollary 4.2]:

(29) $\quad c_0 = 2, \quad c_1 = h, \quad c_2 = \tfrac{1}{2}h^2, \quad c_3 = \tfrac{1}{3}\left(\tfrac{1}{2}h^3 + \tfrac{1}{4}\beta h - \tfrac{1}{2}\gamma\right), \quad c_4 = \tfrac{1}{4}\left(\tfrac{1}{6}h^4 + \tfrac{1}{3}\beta h^2 - \tfrac{1}{3}2\gamma h\right),$

and for each $n \geq 1$,

(30) $\qquad (n+4)c_{n+4} - \tfrac{1}{2}(n+2)\beta c_{n+2} + \left(\tfrac{1}{4}\beta\right)^2 n c_n = h c_{n+3} - \left(\tfrac{1}{4}\beta h + \tfrac{1}{2}\gamma\right)c_{n+1}.$

In order to evaluate the determinant giving $[B_{\boldsymbol{P}}(8)]^{\mathrm{virt}}$, we shall use Proposition 6.4 coupled with the formula of Thaddeus [46] determining all top intersection numbers of tautological classes on $\mathrm{SU}_X(2, \omega(p))$.

Precisely, for $m + 2n + 3p = 3g - 3$, one has

$$(31) \qquad \int_{\mathrm{SU}_X(2,\omega(p))} \alpha^m \cdot \beta^n \cdot \gamma^p = (-1)^{g-p} \frac{g!\,m!}{(g-p)!\,q!} 2^{2g-2-p}(2^q - 2)B_q,$$

where $q = m + p + 1 - g$ and B_q denotes the Bernoulli number; those appearing in our calculation are

$$B_2 = \tfrac{1}{6}, \quad B_4 = -\tfrac{1}{30}, \quad B_6 = \tfrac{1}{42}, \quad B_8 = -\tfrac{1}{30}, \quad B_{10} = \tfrac{5}{66}, \quad B_{12} = -\tfrac{691}{2730},$$
$$B_{14} = \tfrac{7}{6}, \quad B_{16} = -\tfrac{3617}{510}, \quad B_{18} = \tfrac{43867}{798}, \quad B_{20} = -\tfrac{174611}{330}, \quad B_{22} = \tfrac{854513}{138}, \quad B_{24} = -\tfrac{236364091}{2730}.$$

Theorem 6.5 *For a general curve X of genus 13, the locus $\mathrm{SU}_X(2, \omega, 8)$ consists of three reduced points corresponding to stable vector bundles.*

Proof As explained, the Lagrangian degeneracy locus $B_P(8)$ is expected to be a curve and we write

$$[B_P(8)]^{\mathrm{virt}} = f(\alpha, \beta, \gamma) + h \cdot u(\alpha, \beta, \gamma),$$

where $f(\alpha, \beta, \gamma)$ and $u(\alpha, \beta, \gamma)$ are homogeneous polynomials of degrees $36 = 3g - 3$ and $35 = 3g - 4$, respectively.

Observe that if $E \in \mathrm{SU}_X(2, \omega, 8)$ then necessarily E is a stable bundle. Otherwise E is strictly semistable, in which case $E = B \oplus (\omega_X \otimes B^\vee)$, where $B \in W_{12}^3(X)$, which contradicts the Brill–Noether theorem on X. Since ρ is a $\mathbb{P}^1$–fibration over the locus of stable vector bundles, it follows that $B_P(8)$ is a $\mathbb{P}^1$–fibration over $\mathrm{SU}_X(2, \omega, 8)$. Furthermore, applying [45], the Mukai-Petri map μ_E is an isomorphism for each vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$, therefore $\mathrm{SU}_X(2, \omega, 8)$ is a reduced zero-dimensional cycle. We denote by a its length, thus we can write

$$(32) \qquad [B_P(8)] = [B_P(8)]^{\mathrm{virt}} = a\rho^*([E_0]) = f(\alpha, \beta, \gamma) + h \cdot u(\alpha, \beta, \gamma),$$

where $[E_0] \in \mathrm{SU}_X(2, \omega)$ is general. Intersecting both sides of (32) with h, we obtain

$$h \cdot f(\alpha, \beta, \gamma) = -h \cdot \alpha u(\alpha, \beta, \gamma).$$

Next observe that $\rho^*([E_0]) \cdot \alpha = 2$. Indeed, since ρ is a $\mathbb{P}^1$–fibration over the open locus of stable bundles and $\omega_P = \rho^*(\mathcal{L}_{\mathrm{ev}}) \otimes \pi^*(-\alpha)$, it follows that

$$-2 = \deg(\omega_{P|\rho^*([E_0])}) = \omega_P \cdot \rho^*([E_0]) = -\alpha \cdot \rho^*([E_0]).$$

Intersecting both sides of (32) with α, we find $2a = h \cdot \alpha u(\alpha, \beta, \gamma) = -h \cdot f(\alpha, \beta, \gamma)$, so

$$a = |\mathrm{SU}_X(2, \omega, 8)| = \frac{1}{2} \int_P h f(\alpha, \beta, \gamma) = \frac{1}{2} \int_{\mathrm{SU}_X(2,\omega(p))} f(\alpha, \beta, \gamma).$$

We are left with the task of computing the degree 36 polynomial $f(\alpha, \beta, \gamma)$, which is a long but elementary calculation. We consider the determinant (28) computing the class of $B_P(8)$. First we substitute for each of the classes $c_1, \ldots, c_{15}$ the expression in terms of α, β, γ and h given by the recursion (30), starting with the initial conditions (29). Evaluating this determinant, we obtain a polynomial of degree 36 in the classes

α, β, γ and h. We recursively express all the powers h^n with $n \geq 2$ and obtain a formula of the form $[B_P(8)] = f(\alpha, \beta, \gamma) + h \cdot u(\alpha, \beta, \gamma)$. We set $h = 0$ in this formula and then we evaluate each monomial of degree 36 in α, β and γ using Thaddeus' formulas (31). At the end, we obtain $f(\alpha, \beta, \gamma) = -6$, which completes the proof of Theorem 6.5.[4] $\qquad\square$

7 The nonabelian Brill–Noether divisor on $\overline{\mathcal{M}}_{13}$

In this section we determine the class of the nonabelian Brill–Noether divisor $\overline{\mathcal{MP}}_{13}$, and prove Theorem 1.1. The results in this section also lay the groundwork for the proof that $\overline{\mathcal{R}}_{13}$ is of general type.

7.1 Tautological classes on the universal nonabelian Brill–Noether locus

Definition 7.1 Let $\mathfrak{M}_{13}^\sharp$ be the open substack of $\overline{\mathfrak{M}}_{13}$ consisting of

  (i) smooth curves X of genus 13 with $\mathrm{SU}_X(2, \omega, 9) = \varnothing$, or

  (ii) 1–nodal irreducible curves $[X/y \sim q]$, where X is a 7–gonal smooth genus 12 curve, $y, q \in X$, and the multiplication map $\phi_L \colon \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$ is surjective for each $L \in W_{15}^5(X)$.

Let $\mathcal{M}_{13}^\sharp$ be the open subset of $\overline{\mathcal{M}}_{13}$ coarsely representing $\mathfrak{M}_{13}^\sharp$.

Note that $\mathcal{M}_{13}^\sharp$ and $\mathcal{M}_{13} \cup \Delta_0$ agree in codimension one, in particular we identify $CH^1(\mathcal{M}_{13}^\sharp)$ with $\mathbb{Q}\langle \lambda, \delta_0 \rangle$. We let $\mathfrak{SU}_{13}^\sharp(2, \omega, 8)$ be the moduli stack of pairs $[X, E]$, where $[X] \in \mathcal{M}_{13}^\sharp$ and E is a semistable rank-two vector bundle on X with $\det(E) \cong \omega_X$ and $h^0(X, E) \geq 8$. Let $\mathcal{SU}_{13}^\sharp(2, \omega, 8)$ be the coarse moduli space of $\mathfrak{SU}_{13}^\sharp(2, \omega, 8)$. We still denote by $\vartheta \colon \mathfrak{SU}_{13}^\sharp(2, \omega, 8) \to \mathfrak{M}_{13}^\sharp$ the forgetful map.

Proposition 7.2 *The map $\vartheta \colon \mathfrak{SU}_{13}^\sharp(2, \omega, 8) \to \mathfrak{M}_{13}^\sharp$ is proper. Moreover, for each $[X, E] \in \mathcal{SU}_{13}^\sharp(2, \omega, 8)$ the corresponding vector bundle E is globally generated.*

Proof Suppose $\mathcal{X} \to T$ is a flat family of stable curves of genus 13, whose generic fiber X_η is smooth and the special fiber X_0 corresponds to a 1–nodal curve in $\mathcal{M}_{13}^\sharp$. The moduli space $\mathrm{SU}_{X_\eta}(2, \omega)$ specializes to a moduli space $\mathrm{SU}_{X_0}(2, \omega)$ that is a closed subvariety of the moduli space $U_{X_0}(2, 24)$ of S–equivalence classes of torsion-free sheaves of rank-two and degree 24 on X_0. The points in $\mathrm{SU}_{X_0}(2, \omega)$ are described in [42].

We claim that if $E \in \mathrm{SU}_{X_0}(2, \omega)$ satisfies $h^0(X_0, E) \geq 8$, then necessarily E is locally free, in which case $\bigwedge^2 E \cong \omega_{X_0}$. Suppose $\nu \colon X \to X_0$ is the normalization map, let $y, q \in X$ denote the inverse images of the node p of X_0 and assume E is not locally free at p. Denoting by $\mathfrak{m}_p \subseteq \mathcal{O}_{X_0, p}$ the maximal ideal, either

  (i) $E_p \cong \mathfrak{m}_p \oplus \mathfrak{m}_p$, or

  (ii) $E_p \cong \mathcal{O}_{X_0, p} \oplus \mathfrak{m}_p$.

---

[4]The *Maple* file describing all calculations explained here is at https://www.mathematik.hu-berlin.de/farkas/gen13bn.mw.

In the first case, $E = \nu_*(F)$, where F is a vector bundle of rank two on X with $\det(F) \cong \omega_X$, that is, $\mathrm{SU}_X(2, \omega, 8) \neq \varnothing$. Note that

$$h^0(X, \det(F)) = 12 \leq 2h^0(X, F) - 4,$$

implying that F has a subpencil $A \hookrightarrow F$.[5] Then $A \in W_7^1(X)$ and $L := \omega_X \otimes A^\vee \in W_{15}^5(X)$ is such that $\phi_L \colon \mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$ is not surjective. This is ruled out by the definition of $\mathcal{M}_{13}^\sharp$. In case (ii), when $E_p \cong \mathcal{O}_{X_0, p} \oplus \mathfrak{m}_p$, one has an exact sequence

$$0 \to E \to \nu_*(\widetilde{F}) \to K(p) \to 0,$$

where $\widetilde{F} = \nu^*(E)/\mathrm{Torsion}$ is a vector bundle on the smooth curve X and satisfies $\det(F) = \omega_X(y)$, or $\det(F) \cong \omega_X(q)$; see also [42, 1.2]. Observe that also in this case F necessarily carries a subpencil, and we argue as before to rule out this possibility.

We now turn out to the last part of Proposition 7.2. Choose $[X, E] \in \mathcal{SU}_{13}^\sharp(2, \omega, 8)$ and assume for simplicity X is smooth (the case when X is 1–nodal being similar). Assume E is not globally generated at a point $q \in X$. Then there exists a vector bundle $F \in \mathrm{SU}_X(2, \omega(-q), 8)$, obtained from E by an elementary transformation at q. Note that $h^0(X, \det F) \leq 2h^0(X, F) - 4$, which forces F to have a subpencil $A \hookrightarrow F$. Necessarily, $\deg(A) = 7$. Since $h^0(F) = h^0(A) + h^0(\omega_X \otimes A^\vee(-q))$, setting $L := \omega_X \otimes A^\vee \in W_{17}^6(X)$, it follows that the multiplication map

$$H^0(X, L) \otimes H^0(X, L(-q)) \to H^0(X, L^{\otimes 2}(-q))$$

is not surjective, and in particular the map $\mathrm{Sym}^2 H^0(X, L) \to H^0(X, L^{\otimes 2})$ is not surjective either. Then X possesses a stable rank-two vector bundle with canonical determinant and $9 = h^0(X, A) + h^0(X, L)$ sections, which is not the case. $\square$

Let us consider the universal genus 13 curve

$$\wp \colon \mathcal{C}_{13}^\sharp \to \mathcal{SU}_{13}^\sharp(2, \omega, 8),$$

then let $\mathfrak{E}$ be the universal rank-two bundle over the stack $\mathcal{SU}_{13}^\sharp(2, \omega, 8)$. Note that we can normalize $\mathfrak{E}$ in such a way that $\det(\mathfrak{E}) \cong \omega_\wp$.

Definition 7.3 We define the tautological class $\gamma := \wp_*(c_2(\mathfrak{E})) \in CH^1(\mathcal{SU}_{13}^\sharp(2, \omega, 8))$.

We aim to determine the pushforward to $\mathcal{M}_{13}^\sharp$ of the class γ in terms of λ and δ_0. To that end, we begin with the following:

Proposition 7.4 *The pushforward $\wp_*(\mathfrak{E})$ is a locally free sheaf of rank 8 and*

$$c_1(\wp_*(\mathfrak{E})) = \vartheta^*(\lambda) - \tfrac{1}{2}\gamma \in CH^1(\mathcal{SU}_{13}^\sharp(2, \omega, 8)).$$

---

[5] Use that for dimension reasons the determinant map $d \colon \bigwedge^2 H^0(X, F) \to H^0(X, \det(F))$ must necessarily vanish on a pure element $0 \neq s_1 \wedge s_2$, with $s_1, s_2 \in H^0(X, F)$. The subpencil in question is then generated by the sections s_1 and s_2.

Proof The fact that $\wp_*(\mathfrak{E})$ is locally free follows from [29]. We apply Grothendieck–Riemann–Roch to the curve $\wp\colon \mathfrak{C}_{13}^{\sharp} \to \mathfrak{SU}_{13}^{\sharp}(2,\omega,8)$ and to the vector bundle $\mathfrak{E}$ to obtain

$$\mathrm{ch}(\wp_!(\mathfrak{E})) = \wp_*\big[\big(2+c_1(\mathfrak{E})+\tfrac{1}{2}(c_1^2(\mathfrak{E})-2c_2(\mathfrak{E}))+\cdots\big)\cdot\big(1-\tfrac{1}{2}c_1(\Omega_{\wp}^1)+\tfrac{1}{12}(c_1^2(\Omega_{\wp}^1)+c_2(\Omega_{\wp}^1))+\cdots\big)\big].$$

We consider the degree-one terms in this equality. Using [27, page 49], observe that

$$c_1(\Omega_{\wp}^1) = c_1(\omega_{\wp}) \quad \text{and} \quad \wp_*\big(\tfrac{1}{12}(c_1^2(\Omega_{\wp}^1)+c_2(\Omega_{\wp}^1))\big) = \vartheta^*(\lambda).$$

By Serre duality, observe that $R^1\wp_*(\mathfrak{E}) \cong \wp_*(\mathfrak{E})^{\vee}$, therefore one can write

$$2c_1(\wp_*(\mathfrak{E})) = c_1(\wp_*(\mathfrak{E})) - c_1(R^1\wp_*(\mathfrak{E})) = 2\vartheta^*(\lambda) - \tfrac{1}{2}\wp_*(c_1^2(\omega_{\wp})) + \tfrac{1}{2}\wp_*(c_1^2(\omega_{\wp})) - \gamma,$$

which leads to the claimed formula. $\qquad\square$

In view of our future applications to $\overline{\mathcal{R}}_{13}$, we introduce the rank-six vector bundle

$$\mathcal{M}_{\mathfrak{E}} := \mathrm{Ker}\{\wp^*(\wp_*(\mathfrak{E})) \to \mathfrak{E}\}.$$

The fiber $M_E := \mathcal{M}_{\mathfrak{E}}[X,E]$ over a point $[X,E] \in \mathcal{SU}_{13}^{\sharp}(2,\omega,8)$ sits in an exact sequence

$$(33) \qquad\qquad 0 \to M_E \to H^0(X,E)\otimes\mathcal{O}_X \xrightarrow{\mathrm{ev}} E \to 0,$$

where exactness on the right is a consequence of Proposition 7.2.

Proposition 7.5 *The following formulas hold*:

$$c_1(\mathcal{M}_{\mathfrak{E}}) = \wp^*\big(\vartheta^*(\lambda)-\tfrac{1}{2}\gamma\big) - c_1(\omega_{\wp}),$$
$$c_2(\mathcal{M}_{\mathfrak{E}}) = \wp^*c_2(\wp_*\mathfrak{E}) - c_2(\mathfrak{E}) - c_1(\omega_{\wp})\cdot\wp^*\big(\vartheta^*(\lambda)-\tfrac{1}{2}\gamma\big) + c_1^2(\omega_{\wp}).$$

Proof This follows from the splitting principle applied to $\mathcal{M}_{\mathfrak{E}}$, coupled with Proposition 7.4. $\qquad\square$

7.2 The resonance divisor in genus 13

A general curve X of genus 13 has 3 stable vector bundles $E \in \mathrm{SU}_X(2,\omega,8)$. In this case $h^0(X,\det(E)) = 2h^0(X,E)-3$, which implies that requiring E to carry a subpencil defines a divisorial condition on the moduli space $\mathcal{SU}_{13}(2,\omega,8)$ and thus on $\mathcal{M}_{13}$. For a vector bundle $E \in \mathrm{SU}_X(2,\omega)$, we denote its determinant map by

$$d\colon \bigwedge\nolimits^2 H^0(X,E) \to H^0(X,\omega_X).$$

Definition 7.6 The *resonance divisor* $\mathfrak{Res}_{13}^{\sharp}$ is the locus of curves $[X] \in \mathcal{M}_{13}^{\sharp}$ for which

$$G(2,H^0(X,E))\cap\mathbb{P}(\mathrm{Ker}(d)) \neq \varnothing$$

for some vector bundle $E \in \mathrm{SU}_X(2,\omega,8)$. In other words, $\mathfrak{Res}_{13}^{\sharp}$ is the locus of $[X]$ for which there exists an element $0 \neq s_1\wedge s_2 \in \bigwedge^2 H^0(X,E)$ such that $d(s_1\wedge s_2) = 0$.

We set $\mathfrak{Res}_{13} := \mathfrak{Res}^{\sharp}_{13} \cap \mathcal{M}_{13}$. Note that $\mathfrak{Res}^{\sharp}_{13}$ comes with an induced scheme structure under the proper map $\vartheta \colon \mathfrak{SU}^{\sharp}_{13}(2, \omega, 8) \to \mathfrak{M}^{\sharp}_{13}$. The points in $\mathfrak{Res}^{\sharp}_{13}$ correspond to those curves X for which a vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$ carries a subpencil (which is generated by the sections $s_1, s_2 \in H^0(X, E)$ with $d(s_1 \wedge s_2) = 0$). The class $[\mathfrak{Res}^{\sharp}_{13}]$ can be computed in terms of certain tautological classes over $\mathfrak{SU}^{\sharp}_{13}(2, \omega, 8)$. On the other hand, we have a geometric characterization of points in $\mathfrak{Res}_{13}$, and it turns out that the resonance divisor coincides with $\mathfrak{D}_{13}$ away from the heptagonal locus $\mathcal{M}^1_{13,7}$.

Proof of Theorem 1.7 We show that one has the following equality of effective divisors

$$\mathfrak{Res}_{13} = \mathfrak{D}_{13} + 3 \cdot \mathcal{M}^1_{13,7}$$

on $\mathcal{M}_{13}$. Indeed, let us assume that $[X] \in \mathfrak{Res}_{13} \setminus \mathcal{M}^1_{13,7}$, and let $E \in \mathrm{SU}_X(2, \omega, 8)$ be the vector bundle which can be written as an extension

$$(34) \qquad\qquad 0 \to A \to E \to \omega_X \otimes A^{\vee} \to 0,$$

where $h^0(X, A) \geq 2$. Since $\mathrm{gon}(X) = 8$, and since $8 \leq h^0(X, E) \leq h^0(X, A) + h^0(X, \omega_X \otimes A^{\vee})$, it follows that $A \in W^1_8(X)$ and $L := \omega_X \otimes A^{\vee} \in W^5_{16}(X)$. If such an extension exists, then the map ϕ_L is not surjective, therefore $[X] \in \mathfrak{D}_{13}$.

Conversely, if $[X] \in \mathfrak{D}_{13}$, there is some $L \in W^5_{16}(X)$ such that the multiplication map ϕ_L is not surjective. For $[X]$ a general point of an irreducible component of $\mathfrak{D}_{13}$, we may assume that the multiplication map ϕ_L has corank one, for otherwise $\varphi_L \colon X \hookrightarrow \mathbb{P}^5$ lies on a $(2, 2, 2)$ complete intersection in $\mathbb{P}^5$, which is a (possibly degenerate) $K3$ surface. But the locus of curves $[X] \in \mathcal{M}_{13}$ lying on a (possibly degenerate) $K3$ surface cannot exceed $g + 19 = 32 < 3g - 4$, a contradiction. We let

$$E \in \mathbb{P}(\mathrm{Ext}^1(L, \omega_X \otimes L^{\vee}))$$

be the *unique* vector bundle with $h^0(X, E) = h^0(X, L) + h^0(X, \omega_X \otimes L^{\vee}) = 8$. The argument of Proposition 6.2 shows that E is stable, otherwise there would exist an effective divisor M of degree 4 on X such that $L(-M) \in W^3_{12}(X)$. Since $\rho(13, 3, 12) = -3$, the locus of curves $[X] \in \mathcal{M}_{13}$ with $W^3_{12}(X) \neq \varnothing$ has codimension at least three in $\mathcal{M}_{13}$, hence this situation does not occur along a component of $\mathfrak{D}_{13}$. Summarizing, away from the divisor $\mathcal{M}^1_{13,7}$, the divisors $\mathfrak{Res}_{13}$ and $\mathfrak{D}_{13}$ coincide.

We now show that $\mathcal{M}^1_{13,7}$ appears with multiplicity 3 inside $\mathfrak{Res}_{13}$. Let X be a general 7–gonal curve of genus 13 and let $A \in W^1_7(X)$ denote its (unique) degree 7 pencil. Set $L := \omega_X \otimes A^{\vee} \in W^6_{17}(X)$. Each vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$ that has a subpencil appears as an extension

$$(35) \qquad\qquad 0 \to A \to E \xrightarrow{j} L \to 0.$$

In this case $h^0(X, E) = h^0(X, A) + h^0(X, L) - 1$. That is, $V := \mathrm{Im}\{H^0(E) \xrightarrow{j} H^0(L)\}$ is 6–dimensional. Furthermore, the multiplication map

$$\mu_V \colon V \otimes H^0(X, L) \to H^0(X, L^{\otimes 2})$$

is not surjective. Conversely, each 6–dimensional subspace $V \subseteq H^0(X, L)$ such that μ_V is not surjective leads to a vector bundle $E \in \mathbb{P}(\text{Ext}^1(L, A))$ with $h^0(X, E) = 8$. The corresponding bundle E is stable unless V is of the form $H^0(X, L(-p))$ for a point $p \in X$, in which case E can also be realized as an extension

$$0 \to L(-p) \to E \to A(p) \to 0.$$

To determine the number of such subspaces $V \subseteq H^0(X, L)$, we consider the projective space $\boldsymbol{P}^6 := \mathbb{P}(H^0(X, L)^\vee)$ and consider the vector bundle $\mathcal{A}$ on $\boldsymbol{P}^6$ with fiber

$$\mathcal{A}(V) = \frac{V \otimes H^0(X, L)}{\bigwedge^2 V}$$

over a point $[V] \in \boldsymbol{P}^6$. There exists a bundle morphism $\mu : \mathcal{A} \to H^0(X, L^{\otimes 2}) \otimes \mathcal{O}_{\boldsymbol{P}^6}$ given by multiplication and the subspaces $[V] \in \boldsymbol{P}^6$ for which μ_V is not surjective (or, equivalently, $\mu^\vee$ is not injective) are precisely those lying in the degeneracy locus of μ, that is, for which $\text{rk}(\mu(V)) = 21$. Applying the Porteous formula we find

$$[Z_{21}(\mu)] = c_6\big(H^0(X, L^{\otimes 2})^\vee \otimes \mathcal{O}_{\boldsymbol{P}^6} - \mathcal{A}^\vee\big) = c_6(-\mathcal{A}).$$

To compute the Chern classes of $\mathcal{A}$, we recall that via the Euler sequence the rank-six vector bundle $M_{\boldsymbol{P}^6}$ on $\boldsymbol{P}^6$ with $M_{\boldsymbol{P}^6}(V) = V \subseteq H^0(X, L)$ can be identified with $\Omega_{\boldsymbol{P}^6}(1)$. Then $\mathcal{A}$ is isomorphic to $M_{\boldsymbol{P}^6} \otimes H^0(X, L)/\bigwedge^2 M_{\boldsymbol{P}^6}$. From the exact sequence

$$0 \to \bigwedge^2 M_{\boldsymbol{P}^6} \to \bigwedge^2 H^0(X, L) \otimes \mathcal{O}_{\boldsymbol{P}^7} \to M_{\boldsymbol{P}^6}(1) \to 0,$$

recalling that $c_{\text{tot}}(M_{\boldsymbol{P}^6}) = 1/(1 + h)$, where $h = c_1(\mathcal{O}_{\boldsymbol{P}^6}(1))$, we find $c_{\text{tot}}\big(\bigwedge^2 M_{\boldsymbol{P}^6}\big) = (1 + 2h)/(1 + h)^7$, therefore

$$[Z_{21}(\mu)] = \left[\frac{1}{(1 + h)^7} \cdot \frac{(1 + h)^7}{1 + 2h} \right]_6 = \left[\frac{1}{1 + 2h} \right]_6 = 2^6 \cdot h^6 = 64.$$

From this, we subtract the excess contribution corresponding to the locus $X \xrightarrow{|L|} \boldsymbol{P}^6$, parametrizing the subspaces $V = H^0(X, L(-p))$ corresponding to unstable bundles. Via the excess Porteous formula [24, Example 14.4.7], this locus appears in the class $[Z_{21}(\mu)]$ with a contribution of

$$c_1\big(\text{Ker}(\mu^\vee) \otimes \text{Coker}(\mu^\vee) - N_{X/\boldsymbol{P}^6}\big) = -5c_1(\text{Ker}(\mu^\vee)) + c_1(\mathcal{A}_{|X}^\vee) - c_1(N_{X/\boldsymbol{P}^6}).$$

The restriction to $X \subseteq \boldsymbol{P}^6$ of the kernel bundle of $\mu^\vee$ can be identified with $L^\vee$, whereas $c_1(\mathcal{A}_{|X}^\vee) = -2c_1(M_{\boldsymbol{P}^6|X}) = 2 \deg(L)$. Furthermore $c_1(N_{X/\boldsymbol{P}^6}) = 7 \deg(L) + 2g(X) - 2$. All in all, the excess contribution to $[Z_{21}(\mu)]$ coming from X equals

$$10 \deg(L) + 2 \deg(L) - 7 \deg(L) - 2g(X) - 2 = 5 \cdot 17 - 24 = 61.$$

Therefore, for a general curve $[X] \in \mathcal{M}_{13,7}^1$, there are $3 = 64 - 61$ vector bundles $E \in \text{SU}_X(2, \omega, 8)$ having A as a subpencil, which finishes the proof. $\qquad\square$

We are now in a position to explain how Theorems 1.3 and 1.7 provide enough geometric information to determine the pushforward to $\mathfrak{M}_{13}^\sharp$ of the class γ.

Proposition 7.7 *One has* $\vartheta_*(\gamma) = \frac{11288}{143}\lambda - \frac{1582}{143}\delta_0 \in CH^1(\mathcal{M}_{13}^\sharp)$.

Proof The divisor $\mathfrak{Res}_{13}^\sharp$ is defined as the pushforward under $\vartheta \colon \mathfrak{SU}_{13}^\sharp(2, \omega, 8) \to \mathfrak{M}_{13}^\sharp$ of the locus where the fibers of the morphism of vector bundles

$$d \colon \bigwedge\nolimits^2 \wp_*(\mathfrak{E}) \to \wp_*(\omega_\wp)$$

contain a rank-two tensor in their kernel. To compute the class of this locus, we use Proposition 7.4 in combination with [22, Theorem 1.1]:[6]

$$[\mathfrak{Res}_{13}^\sharp] = 132\big(c_1(\wp_*(\omega_\wp)) - \tfrac{13}{4}c_1(\wp_*(\mathfrak{E}))\big) = 132\big(-\tfrac{9}{4}\vartheta^*(\lambda) + \tfrac{13}{8}\gamma\big).$$

Using [27], we write $[\overline{\mathcal{M}}_{13,7}^1] = 6 \cdot (48\lambda - 7\delta_0 - \cdots)$ for the class of the heptagonal locus, while the class $[\widetilde{\mathfrak{D}}_{13}]$ is computed by Theorem 1.4. Since $\deg(\vartheta) = 3$, we then find

$$\vartheta_*(\gamma) = \tfrac{48}{13}\big(\tfrac{5059}{264}\lambda - \tfrac{749}{264}\delta_0 + \tfrac{9}{8}\lambda + \tfrac{3}{132}(48\lambda - 7\delta_0)\big) = \tfrac{1128}{143}\lambda - \tfrac{1582}{143}\delta_0. \qquad \square$$

7.3 The class of the nonabelian Brill–Noether divisor on $\overline{\mathcal{M}}_{13}$

In the introduction, we defined the nonabelian Brill–Noether divisor $\mathcal{MP}_{13}^\sharp$ as the locus of curves $[X] \in \mathcal{M}_{13}^\sharp$ for which there exists $E \in \mathrm{SU}_X(2, \omega, 8)$ such that the map

$$\mu_E \colon \mathrm{Sym}^2 H^0(X, E) \to H^0(X, \mathrm{Sym}^2 E)$$

is not an isomorphism, or equivalently, the scheme $\mathrm{SU}_X(2, \omega, 8)$ is not reduced. We now compute the class of this divisor.

Proof of Theorem 1.1 The locus $\mathcal{MP}_{13}^\sharp$ is the pushforward under the proper map ϑ of the degeneracy locus of the following map of vector bundles over $\mathfrak{SU}_{13}^\sharp(2, \omega, 8)$:

$$\mathrm{Sym}^2 \wp_*(\mathfrak{E}) \to \wp_*(\mathrm{Sym}^2 \mathfrak{E}).$$

Using Grothendieck–Riemann–Roch for $\wp \colon \mathfrak{C}_{13}^\sharp \to \mathfrak{SU}_{13}^\sharp(2, \omega, 8)$, we compute

$$c_1(p_*(\mathrm{Sym}^2 \mathfrak{E})) = \wp_*\big[\big(3 + 3c_1(\mathfrak{E}) + \tfrac{1}{2}(5c_1^2(\mathfrak{E}) - 8c_2(\mathfrak{E}))\big) \cdot \big(1 - \tfrac{1}{2}c_1(\Omega_\wp^1) + \tfrac{1}{12}(c_1^2(\Omega_\wp^1) + c_2(\Omega_\wp^1))\big)\big]_2.$$

Using again that $12\wp_*\big(c_1^2(\Omega_\wp^1) + c_2(\Omega_\wp^1)\big) = \vartheta^*(\lambda)$, we conclude that

$$c_1(\wp_*(\mathrm{Sym}^2 \mathfrak{E})) = 3\vartheta^*(\lambda) + \wp_*(c_1^2(\omega_\wp)) - 4\gamma = \vartheta^*(15\lambda - \delta_0) - 4\gamma.$$

Via Proposition 7.4, we have $c_1(\mathrm{Sym}^2 \wp_*(\mathfrak{E})) = 9c_1(\wp_*(\mathfrak{E})) = 9\big(\vartheta^*(\lambda) - \tfrac{1}{2}\gamma\big)$, yielding

$$[\mathcal{MP}_{13}^\sharp] = \vartheta_*\big(c_1(\wp_*(\mathrm{Sym}^2 \mathfrak{E}) - \mathrm{Sym}^2 \wp_*(\mathfrak{E}))\big) = 3(6\lambda - \delta_0) + \tfrac{1}{2}\vartheta_*(\gamma).$$

Substituting via Proposition 7.7, we find $[\mathcal{MP}_{13}^\sharp] = \tfrac{1}{143}(8218\lambda - 1220\delta_0)$. $\qquad \square$

---

[6]The result in [22] is stated for a morphism of vector bundles of the form $\mathrm{Sym}^2(\mathcal{E}) \to \mathcal{F}$. An immediate inspection of the proof shows though that the *same formula* applies also in the setting of a morphism of the form $\bigwedge^2(\mathcal{E}) \to \mathcal{F}$.

8 The Kodaira dimension of $\overline{\mathcal{R}}_{13}$

We turn our attention to showing that the Prym moduli space $\overline{\mathcal{R}}_{13}$ is a variety of general type. We begin by recalling basics on the geometry of the moduli of Prym variety, referring to [20] for details. We denote by $\mathfrak{R}_g := \overline{\mathcal{M}}_g(\mathcal{B}\mathbb{Z}_2)$ the Deligne–Mumford stack of *Prym curves* of genus g classifying triples $[Y, \eta, \beta]$, where Y is a nodal curve of genus g such that each of its rational components meets the rest of the curve in at least two points, $\eta \in \mathrm{Pic}^0(Y)$ is a line bundle of total degree 0 such that $\eta_{|R} = \mathcal{O}_R(1)$ for every rational component $R \subseteq Y$ with $|R \cap \overline{Y \setminus R}| = 2$ (such a component is called *exceptional*), and $\beta : \eta^{\otimes 2} \to \mathcal{O}_Y$ is a morphism which is generically nonzero along each nonexceptional component of Y. Let $\overline{\mathcal{R}}_g$ be the coarse moduli space of $\mathfrak{R}_g$. One has a finite cover

$$\pi : \overline{\mathcal{R}}_g \to \overline{\mathcal{M}}_g.$$

8.1 The boundary divisors of $\overline{\mathcal{R}}_g$

The geometry of the boundary of $\overline{\mathcal{R}}_g$ is described in [20]; we recall some facts. If $[X_{yq} = X/y \sim q]$ in $\Delta_0 \subseteq \overline{\mathcal{M}}_g$ is such that $[X, y, q] \in \mathcal{M}_{g-1,2}$, denoting by $\nu : X \to X_{yq}$ the normalization map, there are three types of Prym curves in the fiber $\pi^{-1}([X_{yq}])$. First, one can choose a nontrivial 2–torsion point $\eta \in \mathrm{Pic}^0(X_{yq})$. If $\nu^*(\eta) \neq \mathcal{O}_X$, this amounts to choosing a 2–torsion point $\eta_X \in \mathrm{Pic}^0(X)[2] \setminus \{\mathcal{O}_X\}$ together with an identification of the fibers $\eta_X(y)$ and $\eta_X(q)$ at the points y and q, respectively. As we vary $[X, y, q]$, points of this type fill up the boundary divisor Δ'_0 in $\overline{\mathcal{R}}_g$. The Prym curves corresponding to the situation $\nu^*(\eta) \cong \mathcal{O}_X$ fill up the boundary divisor Δ''_0. Finally, choosing a line bundle η_X on X with $\eta_X^{\otimes 2} \cong \mathcal{O}_X(-y-q)$ leads to a Prym curve $[Y := X \cup_{y,q} R, \eta, \beta]$, where R is a smooth rational curve meeting X at y and q, and $\eta \in \mathrm{Pic}^0(Y)$ is a line bundle such that $\eta_{|X} = \eta_X$ and $\eta_{|R} = \mathcal{O}_R(1)$. Points of this type fill up the boundary divisor Δ_0^{ram} of $\overline{\mathcal{R}}_g$, which is the ramification divisor of the morphism π.

Denoting by $\delta'_0 := [\Delta'_0]$, $\delta''_0 := [\Delta''_0]$ and $\delta_0^{\mathrm{ram}} := [\delta_0^{\mathrm{ram}}]$ the corresponding divisor classes, one has the following relation in $CH^1(\overline{\mathcal{R}}_g) \cong CH^1(\overline{\mathfrak{R}}_g)$, see [20]:

$$\pi^*(\delta_0) = \delta'_0 + \delta''_0 + 2\delta_0^{\mathrm{ram}}.$$

The finite morphism $\pi : \overline{\mathcal{R}}_g \to \overline{\mathcal{M}}_g$ being ramified only along the divisor Δ_0^{ram}, one has

$$(36) \qquad K_{\overline{\mathcal{R}}_g} = 13\lambda - 2(\delta'_0 + \delta''_0) - 3\delta_0^{\mathrm{ram}} - 2\sum_{i=1}^{\lfloor g/2 \rfloor}(\delta_i + \delta_{g-i} + \delta_{i:g-i}) - (\delta_1 + \delta_{g-1} + \delta_{1:g-1}),$$

where $\pi^*(\delta_i) = \delta_i + \delta_{g-i} + \delta_{i:g-i}$; see [20, Theorem 1.5] for details.

8.2 The universal theta divisor on $\overline{\mathcal{R}}_{13}$

For a semistable vector bundle $E \in \mathrm{SU}_X(2, \omega)$ on a smooth curve X of genus g, its *Raynaud theta divisor* $\Theta_E := \{\xi \in \mathrm{Pic}^0(X) : H^0(X, E \otimes \xi) \neq 0\}$ is a 2θ–divisor inside the Jacobian of X; see [41].

Definition 8.1 The universal theta divisor Θ_{13} on $\mathcal{R}_{13}$ is defined as the locus of smooth Prym curves $[X, \eta] \in \mathcal{R}_{13}$ for which there exists a vector bundle $E \in \mathrm{SU}_X(2, \omega, 8)$ such that $H^0(X, E \otimes \eta) \neq 0$.

We first show that, as expected, this definition gives rise to a divisor on $\mathcal{R}_{13}$.

Proposition 8.2 *For a general Prym curve $[X, \eta] \in \mathcal{R}_{13}$, one has $H^0(X, E \otimes \eta) = 0$ for all vector bundles $E \in \mathrm{SU}_X(2, \omega, 8)$. It follows that Θ_{13} is an effective divisor on $\mathcal{R}_{13}$.*

Proof Consider the subvariety of $\mathcal{R}_{13} \times_{\mathcal{M}_{13}} \mathcal{SU}_{13}(2, \omega, 8)$ given by

$$\mathcal{Z} := \{[X, \eta, E] : H^0(X, E \otimes \eta) \neq 0\}.$$

Assume for contradiction that $\mathcal{Z}$ surjects onto $\mathcal{R}_{13}$. Then $\mathcal{Z}$ is a union of *irreducible* components of $\mathcal{R}_{13} \times_{\mathcal{M}_{13}} \mathcal{SU}_{13}(2, \omega, 8)$. In particular, $\mathcal{Z}$ surjects onto the irreducible variety $\mathcal{SU}_{13}(2, \omega, 8)$; see Corollary 6.3. Therefore, for every pair $[X, E] \in \mathcal{SU}_{13}(2, \omega, 8)$, there exists a 2–torsion point η on X with $H^0(X, E \otimes \eta) \neq 0$.

We now specialize to the case when E is a strictly semistable vector bundle of the type

$$E = A^{\otimes 3} \oplus (\omega_X \otimes A^{\otimes(-3)}),$$

where $[X, A]$ is a general tetragonal curve of genus 13. Note that $h^0(X, A^{\otimes 3}) = 4$, by [13, Proposition 2.1]. In particular, $h^0(X, E) = 8$. Using [8] the space $\mathcal{R}_{13} \times_{\mathcal{M}_{13}} \mathcal{M}^1_{13,4}$ parametrizing Prym curves over tetragonal curves of genus 13 is irreducible, therefore $H^0(X, A^{\otimes 3} \otimes \eta) \neq 0$ for *every* triple $[X, \eta, A] \in \mathcal{R}_{13} \times_{\mathcal{M}_{13}} \mathcal{M}^1_{13,4}$. We now further specialize the tetragonal curve X to a hyperelliptic curve and $A = A_0(x + y)$, where $A_0 \in W^1_2(X)$ and $x, y \in X$ are general points, whereas

$$\eta = \mathcal{O}_X(p_1 + p_2 + p_3 + p_4 - q_1 - q_2 - q_3 - q_4) \in \mathrm{Pic}^0(X)[2],$$

where $p_1, \ldots, p_4, q_1, \ldots, q_4$ are mutually distinct Weierstrass points of X. It immediately follows that for these choices $H^0(X, A^{\otimes 3} \otimes \eta) = 0$, which is a contradiction. $\square$

We consider the open substack $\mathfrak{R}^\sharp_{13} := \pi^{-1}(\mathfrak{M}^\sharp_{13})$ of $\overline{\mathfrak{R}}_{13}$ and let $\mathcal{R}^\sharp_{13}$ be its associated coarse moduli space. We identify $CH^1(\mathcal{R}^\sharp_{13})$ with the space $\mathbb{Q}\langle \lambda, \delta'_0, \delta''_0, \delta^{\mathrm{ram}}_0 \rangle$. In what follows we extend the structure on the universal theta divisor Θ_{13} to $\mathcal{R}^\sharp_{13}$ and realize it as the pushforward of the degeneracy locus of a map of vector bundles of the same rank over the fiber product

$$\mathfrak{RSU}^\sharp_{13}(2, \omega, 8) := \mathfrak{R}^\sharp_{13} \times_{\mathfrak{m}^\sharp_{13}} \mathfrak{SU}^\sharp_{13}(2, \omega, 8).$$

We start with a triple $[X, \eta, E] \in \mathfrak{RSU}^\sharp_{13}(2, \omega, 8)$. Via Proposition 7.2 the vector bundle E is globally generated and we let $M_E := \mathrm{Ker}\{H^0(X, E) \otimes \mathcal{O}_X \to E\}$. By tensoring with η and taking cohomology in the exact sequence (33), we observe that $H^0(X, E \otimes \eta) \neq 0$ if and only if the coboundary map

$$(37) \qquad \upsilon \colon H^1(X, M_E \otimes \eta) \to H^0(X, E) \otimes H^0(X, \omega_X \otimes \eta)^\vee$$

is not injective. Since clearly $H^0(X, M_E \otimes \eta) = 0$, it follows that

$$h^1(X, M_E \otimes \eta) = -\deg(M_E) + 6(g-1) = 96 = 8 \cdot 12 = h^0(X, E) \cdot h^0(X, \omega_X \otimes \eta).$$

That is, υ is a map between vector space of the same dimension.

By slightly abusing notation, we still denote by

$$\wp : \mathfrak{RC}^{\sharp}_{13} \to \mathfrak{RSU}^{\sharp}_{13}(2, \omega, 8)$$

the universal curve of genus 13 over $\mathfrak{RSU}^{\sharp}_{13}(2, \omega, 8)$. It comes equipped with a universal rank-two vector bundle $\mathfrak{E}$ such that $\bigwedge^2 \mathfrak{E} \cong \omega_{\wp}$ and $\wp_*(\mathfrak{E})$ is locally free of rank 8 (cf Proposition 7.4), as well as with a universal Prym line bundle $\mathcal{L}$ with $\mathcal{L}_{|\wp^{-1}([X,\eta,E])} \cong \eta$ for any point $[X, \eta, E] \in \mathfrak{RSU}^{\sharp}_{13}(2, \omega, 8)$.

We consider the rank-six vector bundle $\mathcal{M}_{\mathfrak{E}}$ on $\mathcal{RC}^{\sharp}_{13}$ defined by the exact sequence

$$0 \to \mathcal{M}_{\mathfrak{E}} \to \wp^*(\wp_*\mathfrak{E}) \to \mathfrak{E} \to 0,$$

then introduce the following sheaves over $\mathfrak{RSU}^{\sharp}_{13}(2, \omega, 8)$:

$$\mathcal{A} := R^1\wp_*(\mathcal{M}_{\mathfrak{E}} \otimes \mathcal{L}) \quad \text{and} \quad \mathcal{B} := \wp_*(\mathfrak{E}) \otimes \wp_*(\omega_{\wp} \otimes \mathcal{L})^{\vee}.$$

Using the fact that the map υ defined in (37) is a morphism between two vector spaces of the same dimension for every point $[X, \eta, E] \in \mathfrak{RSU}^{\sharp}_{13}(2, \omega, 8)$, via Grauert's theorem we conclude that both $\mathcal{A}$ and $\mathcal{B}$ are locally free of the same rank 96, and there exists a morphism

$$(38) \qquad\qquad\qquad\qquad\qquad \upsilon : \mathcal{A} \to \mathcal{B}$$

whose fiber restrictions are the maps (37). Recall that the forgetful map $\vartheta : \mathfrak{RSU}^{\sharp}_{13}(2, \omega, 8) \to \mathfrak{R}^{\sharp}_{13}$ is generically finite of degree 3. We denote by $\Theta^{\sharp}_{13}$ the pushforward to $\mathcal{R}^{\sharp}_{13}$ of the degeneracy locus of the morphism υ given by (38). Observe that $\Theta^{\sharp}_{13} \cap \mathcal{M}_{13} = \Theta_{13}$.

Theorem 8.3 *The class of the universal theta divisor $\Theta^{\sharp}_{13}$ on $\mathcal{R}_{13}$ is given by*

$$[\Theta^{\sharp}_{13}] = \tfrac{1}{143}\big(10430\lambda - 1582(\delta'_0 + \delta''_0) - \tfrac{5899}{2}\delta^{\mathrm{ram}}_0\big) \in CH^1(\mathcal{R}^{\sharp}_{13}).$$

Proof From Proposition 8.2 it follows that υ is generically nondegenerate, therefore

$$[\Theta^{\sharp}_{13}] = c_1(\mathcal{B} - \mathcal{A}).$$

Computing the class $c_1(\mathcal{B})$ is straightforward. We find that $c_1(\wp_*(\omega_{\wp} \otimes \mathcal{L})) = \vartheta^*\big(\lambda - \tfrac{1}{4}\delta^{\mathrm{ram}}_0\big)$, using [20, Proposition 1.7]. Then via Proposition 7.4, we compute

$$c_1(\mathcal{B}) = 12c_1(\wp_*\mathfrak{E}) - 8c_1(\wp_*(\omega_{\wp} \otimes \mathcal{L})) = 12\big(\vartheta^*(\lambda) - \tfrac{1}{2}\gamma\big) - 8\big(\vartheta^*\big(\lambda - \tfrac{1}{4}\delta^{\mathrm{ram}}_0\big)\big) = \vartheta^*(4\lambda + 2\delta^{\mathrm{ram}}_0) - 6\gamma.$$

To determine $c_1(\mathcal{A})$ we apply Grothendieck–Riemann–Roch to the morphism $\wp$:

$$(39) \quad \mathrm{ch}(\wp_!(\mathcal{M}_{\mathfrak{E}} \otimes \mathcal{L})) = \wp_*\big[\big(6 + c_1(\mathcal{M}_{\mathfrak{E}} \otimes \mathcal{L}) + \tfrac{1}{2}(c_1^2(\mathcal{M}_{\mathfrak{E}} \otimes \mathcal{L}) - 2c_2(\mathcal{M}_{\mathfrak{E}} \otimes \mathcal{L})) + \cdots\big)$$
$$\cdot \big(1 - \tfrac{1}{2}c_1(\Omega^1_{\wp}) + \tfrac{1}{12}(c_1^2(\Omega^1_{\wp}) + c_2(\Omega^1_{\wp})) + \cdots\big)\big].$$

Observe by direct calculation that the formulas

$$c_1(\mathcal{M}_\mathfrak{E} \otimes \mathcal{L}) = c_1(\mathcal{M}_\mathfrak{E}) + 6c_1(\mathfrak{L}) \quad \text{and} \quad c_2(\mathcal{M}_\mathfrak{E} \otimes \mathcal{L}) = c_2(\mathcal{M}_\mathfrak{E}) + 5c_1(\mathcal{M}_\mathfrak{E}) \cdot c_1(\mathcal{L}) + 15c_1^2(\mathcal{L})$$

hold, therefore

$$\wp_*\big(\tfrac{1}{2}(c_1^2(\mathcal{M}_\mathfrak{E} \otimes \mathcal{L}) - 2c_2(\mathcal{M}_\mathfrak{E} \otimes \mathcal{L}))\big) = \wp_*\big(\tfrac{1}{2}(c_1^2(\mathcal{M}_\mathfrak{E}) - 2c_2(\mathcal{M}_\mathfrak{E})) + c_1(\mathcal{M}_\mathfrak{E}) \cdot c_1(\mathcal{L}) + 3c_1^2(\mathcal{L})\big)$$

$$= \gamma - \tfrac{1}{2}\wp_*(c_1^2(\omega_\wp)) = \gamma - \tfrac{1}{2}(\vartheta^*(12\lambda - \delta_0' - \delta_0'' - 2\delta_0^{\mathrm{ram}})),$$

where in the last formula we have used Proposition 7.5, Mumford's formula [27] for the class $\wp_*(c_1^2(\omega_\wp))$, and $2\wp_*(c_1^2(\mathcal{L})) = -\vartheta^*(\delta_0^{\mathrm{ram}})$; see [20, Proposition 1.6].

Substituting in the equation (39), coupled with Proposition 7.5 and also using that via the push–pull formula one has $\wp_*\big(\wp^*\big(\vartheta^*(\lambda) - \tfrac{1}{2}\gamma\big) \cdot c_1(\omega_\wp)\big) = (g-1) \cdot \big(\vartheta^*(\lambda) - \tfrac{1}{2}\gamma\big)$, we obtain

$$c_1(\mathcal{A}) = -7\gamma + \vartheta^*\big(6\lambda + \tfrac{3}{2}\delta_0^{\mathrm{ram}}\big).$$

Putting everything together we find

$$[\Theta_{13}^\sharp] = \vartheta_* c_1(\mathcal{B} - \mathcal{A}) = \vartheta_*\big(\gamma - 2\lambda + \tfrac{1}{2}\delta_0^{\mathrm{ram}}\big) = 2\vartheta_*(\gamma) - 6\lambda + \tfrac{3}{2}\delta_0^{\mathrm{ram}}.$$

Finally, Proposition 7.7 gives $143\vartheta_*(\gamma) = 11288\lambda - 1582(\delta_0' + \delta_0'' + 2\delta_0^{\mathrm{ram}})$ and the conclusion follows. □

We can now complete the proof that $\overline{\mathcal{R}}_{13}$ is of general type.

Proof of Theorem 1.2 It is shown in [20, Theorem 6.1] that any g pluricanonical forms defined on $\overline{\mathcal{R}}_g$ automatically extend to any resolution of singularities, therefore $\overline{\mathcal{R}}_g$ is of general type if and only if the canonical class $K_{\overline{\mathcal{R}}_g}$ is big, that is, it can be expressed as a positive rational combination of an ample and an effective class on $\overline{\mathcal{R}}_g$. To that end we shall use, in addition to the closure $\overline{\Theta}_{13}$ in $\overline{\mathcal{R}}_{13}$ of the universal theta divisor Θ_{13}, the divisor $D_{13:2}$ on $\mathcal{R}_{13}$ consisting of pairs $[X, \eta]$ where the 2–torsion point η lies in the divisorial *difference variety*

$$X_6 - X_6 = \{\mathcal{O}_X(D - E) : D, E \in X_6\} \subseteq \mathrm{Pic}^0(X).$$

It is shown in [20, Theorem 0.2] that up to a positive rational constant, the closure of $D_{13:2}$ inside $\overline{\mathcal{R}}_{13}$ is given by $[\overline{D}_{13:2}] = 19\lambda - 3(\delta_0' + \delta_0'') - \tfrac{13}{4}\delta_0^{\mathrm{ram}} - \cdots \in CH^1(\overline{\mathcal{R}}_{13})$. Observe that by construction, $\Theta_{13}^\sharp$ differs from the restriction of $\overline{\Theta}_{13}$ to $\overline{\mathcal{M}}_{13}^\sharp$ by a (possibly empty) *effective* combination of the divisors Δ_0', Δ_0'' and Δ_0^{ram}; hence, using Theorem 8.3 we can write

$$[\overline{\Theta}_{13}] = \tfrac{1}{143}(10430\lambda - b_0'\delta_0' - b_0''\delta_0'' - b_0^{\mathrm{ram}}\delta_0^{\mathrm{ram}} - \cdots) \in CH^1(\overline{\mathcal{R}}_{13}),$$

where $b_0' \geq 1582$, $b_0'' \geq 1582$ and $b_0^{\mathrm{ram}} \geq \tfrac{5899}{2}$. We consider the effective divisor, on $\overline{\mathcal{R}}_{13}$,

$$\mathcal{D} := \tfrac{65}{674}[\overline{\Theta}_{13}] + \tfrac{1153}{3707}[\overline{D}_{13:2}] = a\lambda - a_0'\delta_0' - a_0''\delta_0'' - a_0^{\mathrm{ram}}\delta_0^{\mathrm{ram}} - \sum_{i=1}^{12} a_i\delta_i - \sum_{i=1}^{6} a_{i,13-i}\delta_{i:13-i},$$

where $a = \tfrac{4362}{337}$, $a_0' \geq 2$, $a_0'' \geq 2$ and $a_0^{\mathrm{ram}} \geq 3$. By an argument using pencils on $K3$ surfaces, one can show that each of the coefficients $a_1, \ldots, a_{12}$ or $a_{1,12}, \ldots, a_{6,7}$ is at least equal to 3. Indeed, each boundary divisor Δ_i or $\Delta_{i:13-i}$ of $\overline{\mathcal{R}}_{13}$ is covered by pencils of reducible Prym curves consisting of two

components, of which one moves in a suitable Lefschetz pencil on a *fixed K*3 surface. The intersection numbers of these pencils with the generators of $CH^1(\overline{\mathcal{R}}_g)$ were computed in [20, Proposition 1.8]. Since $\mathcal{D}$ is the closure in $\overline{\mathcal{R}}_{13}$ of an effective divisor on $\mathcal{R}_{13}$, the intersection number of each such pencil with $\mathcal{D}$ is nonnegative. For instance, for $1 \le i \le 6$ we obtain, in this way, the inequality

$$a_{13-i} \ge a_0'(6i + 18) - a(i + 1) \ge 2(6i + 18) - \tfrac{4362}{337}(i + 1) \ge 3.$$

The inequalities for the remaining coefficients of $\mathcal{D}$ can be handled similarly; see also [20, Proposition 1.9]. Since $a = 12.943\ldots < 13$, comparing the class of $\mathcal{D}$ to that of $K_{\overline{\mathcal{R}}_{13}}$ given in (36), we conclude that $K_{\overline{\mathcal{R}}_{13}}$ can be written as a positive combination of $[\mathcal{D}]$ and a multiple of λ, hence it is big. $\qquad\square$

8.3 The Kodaira dimension of $\overline{\mathcal{M}}_{13,n}$

We indicate how our results on divisors on $\overline{\mathcal{M}}_{13}$ can be used to determine the Kodaira dimension of the moduli space $\overline{\mathcal{M}}_{13,n}$.

Proof of Theorem 1.6 It suffices to show that $\overline{\mathcal{M}}_{13,9}$ is of general type to conclude that the same holds for $\overline{\mathcal{M}}_{13,n}$ when $n \ge 10$. We use the divisor $\mathcal{D}_{13:2^4,1^5}$ considered by Logan [36] and defined as the $\mathfrak{S}_9$–orbit (under the action permuting the marked points) of the locus of pointed curves $[X, p_1, \ldots, p_9] \in \mathcal{M}_{13,9}$ such that

$$h^0(X, \mathcal{O}_X(2p_1 + \cdots + 2p_4 + p_5 + \cdots + p_9)) \ge 2.$$

Up to a positive constant the class of the closure in $\overline{\mathcal{M}}_{13,9}$ of $D_{13:2^4,1^5}$ equals

$$[\overline{\mathcal{D}}_{13:2^4,1^5}] = -\lambda + \tfrac{17}{9} \sum_{i=1}^{9} \psi_i - \tfrac{25}{6}\delta_{0:2} - \cdots \in CH^1(\overline{\mathcal{M}}_{13,9}).$$

(See [17] or [36] for the standard notation on the generators of $CH^1(\overline{\mathcal{M}}_{g,n})$.) If $\pi \colon \overline{\mathcal{M}}_{13,9} \to \overline{\mathcal{M}}_{13}$ is the map forgetting the marked points, a routine calculation shows that the canonical class $K_{\overline{\mathcal{M}}_{13,9}}$ can be expressed as a positive linear combination of $[\overline{\mathcal{D}}_{13:2^4,1^5}]$ and $\pi^*([D])$, where $D \in \mathrm{Eff}(\overline{\mathcal{M}}_{13})$ if and only if $2s(D) - \tfrac{9}{17} < 13$. Observe that the class of the nonabelian Brill–Noether divisor $[\overline{\mathcal{MP}}_{13}]$ verifies this inequality, and the result follows. $\qquad\square$

References

[1] **D Agostini**, **I Barros**, *Pencils on surfaces with normal crossings and the Kodaira dimension of $\overline{\mathcal{M}}_{g,n}$*, Forum Math. Sigma 9 (2021) art. id. e31 MR Zbl

[2] **Y An**, **M Baker**, **G Kuperberg**, **F Shokrieh**, *Canonical representatives for divisor classes on tropical curves and the matrix-tree theorem*, Forum Math. Sigma 2 (2014) art. id. e24 MR Zbl

[3] **M Aprodu**, **G Farkas**, *Koszul cohomology and applications to moduli*, from "Grassmannians, moduli spaces and vector bundles" (D A Ellwood, E Previato, editors), Clay Math. Proc. 14, Amer. Math. Soc., Providence, RI (2011) 25–50 MR Zbl

[4] **M Aprodu**, **G Farkas**, **Ş Papadima**, **C Raicu**, **J Weyman**, *Topological invariants of groups and Koszul modules*, Duke Math. J. 171 (2022) 2013–2046 MR Zbl

[5] **E Arbarello**, **M Cornalba**, **P A Griffiths**, **J Harris**, *Geometry of algebraic curves, I*, Grundl. Math. Wissen. 267, Springer (1985) MR Zbl

[6] **A Beauville**, *Prym varieties and the Schottky problem*, Invent. Math. 41 (1977) 149–196 MR Zbl

[7] **A Bertram**, **B Feinberg**, *On stable rank two bundles with canonical determinant and many sections*, from "Algebraic geometry" (P E Newstead, editor), Lecture Notes in Pure and Appl. Math. 200, Dekker, New York (1998) 259–269 MR Zbl

[8] **R Biggers**, **M Fried**, *Irreducibility of moduli spaces of cyclic unramified covers of genus g curves*, Trans. Amer. Math. Soc. 295 (1986) 59–70 MR Zbl

[9] **G Bruns**, $\overline{\mathcal{R}}_{15}$ *is of general type*, Algebra Number Theory 10 (2016) 1949–1964 MR Zbl

[10] **D Chen**, **G Farkas**, **I Morrison**, *Effective divisors on moduli spaces of curves and abelian varieties*, from "A celebration of algebraic geometry" (B Hassett, J McKernan, J Starr, R Vakil, editors), Clay Math. Proc. 18, Amer. Math. Soc., Providence, RI (2013) 131–169 MR Zbl

[11] **A Chiodo**, **D Eisenbud**, **G Farkas**, **F-O Schreyer**, *Syzygies of torsion bundles and the geometry of the level ℓ modular variety over $\overline{\mathcal{M}}_g$*, Invent. Math. 194 (2013) 73–118 MR Zbl

[12] **F Cools**, **J Draisma**, **S Payne**, **E Robeva**, *A tropical proof of the Brill–Noether theorem*, Adv. Math. 230 (2012) 759–776 MR Zbl

[13] **M Coppens**, **G Martens**, *Linear series on a general k–gonal curve*, Abh. Math. Sem. Univ. Hamburg 69 (1999) 347–371 MR Zbl

[14] **E Cotterill**, **A Alonso Gonzalo**, **N Zhang**, *The strong maximal rank conjecture and higher rank Brill–Noether theory*, J. Lond. Math. Soc. 104 (2021) 169–205 MR Zbl

[15] **D Eisenbud**, **J Harris**, *Limit linear series: basic theory*, Invent. Math. 85 (1986) 337–371 MR Zbl

[16] **D Eisenbud**, **J Harris**, *The Kodaira dimension of the moduli space of curves of genus ≥ 23*, Invent. Math. 90 (1987) 359–387 MR Zbl

[17] **G Farkas**, *Koszul divisors on moduli spaces of curves*, Amer. J. Math. 131 (2009) 819–867 MR Zbl

[18] **G Farkas**, *Brill–Noether with ramification at unassigned points*, J. Pure Appl. Algebra 217 (2013) 1838–1843 MR Zbl

[19] **G Farkas**, **D Jensen**, **S Payne**, *The Kodaira dimensions of $\overline{\mathcal{M}}_{22}$ and $\overline{\mathcal{M}}_{23}$*, preprint (2020) arXiv 2005.00622v2

[20] **G Farkas**, **K Ludwig**, *The Kodaira dimension of the moduli space of Prym varieties*, J. Eur. Math. Soc. 12 (2010) 755–795 MR Zbl

[21] **G Farkas**, **M Popa**, *Effective divisors on $\overline{\mathcal{M}}_g$, curves on $K3$ surfaces, and the slope conjecture*, J. Algebraic Geom. 14 (2005) 241–267 MR Zbl

[22] **G Farkas**, **R Rimányi**, *Quadric rank loci on moduli of curves and $K3$ surfaces*, Ann. Sci. Éc. Norm. Supér. 53 (2020) 945–992 MR Zbl

[23] **G Farkas**, **A Verra**, *Prym varieties and moduli of polarized Nikulin surfaces*, Adv. Math. 290 (2016) 314–328 MR Zbl

[24] **W Fulton**, *Intersection theory*, Ergebnisse der Math. (3) 2, Springer (1984) MR Zbl

[25] **C Haase**, **G Musiker**, **J Yu**, *Linear systems on tropical curves*, Math. Z. 270 (2012) 1111–1140 MR Zbl

[26] **J Harris**, **I Morrison**, *Slopes of effective divisors on the moduli space of stable curves*, Invent. Math. 99 (1990) 321–355 MR Zbl

[27] **J Harris**, **D Mumford**, *On the Kodaira dimension of the moduli space of curves*, Invent. Math. 67 (1982) 23–88 MR Zbl

[28] **J Harris**, **L Tu**, *Chern numbers of kernel and cokernel bundles*, Invent. Math. 75 (1984) 467–475 MR Zbl

[29] **R Hartshorne**, *Stable reflexive sheaves*, Math. Ann. 254 (1980) 121–176 MR Zbl

[30] **D Jensen**, **S Payne**, *Tropical independence, I: Shapes of divisors and a proof of the Gieseker–Petri theorem*, Algebra Number Theory 8 (2014) 2043–2066 MR Zbl

[31] **D Jensen**, **S Payne**, *Tropical independence, II: The maximal rank conjecture for quadrics*, Algebra Number Theory 10 (2016) 1601–1640 MR Zbl

[32] **D Khosla**, *Tautological classes on moduli spaces of curves with linear series and a push-forward formula when $\rho = 0$*, preprint (2007) arXiv 0704.1340

[33] **H Lange**, **M S Narasimhan**, *Maximal subbundles of rank two vector bundles on curves*, Math. Ann. 266 (1983) 55–72 MR Zbl

[34] **H Lange**, **P E Newstead**, **S S Park**, *Nonemptiness of Brill–Noether loci in $M(2, K)$*, Comm. Algebra 44 (2016) 746–767 MR Zbl

[35] **R Lazarsfeld**, *Brill–Noether–Petri without degenerations*, J. Differential Geom. 23 (1986) 299–307 MR Zbl

[36] **A Logan**, *The Kodaira dimension of moduli spaces of curves with marked points*, Amer. J. Math. 125 (2003) 105–138 MR Zbl

[37] **S Mukai**, *Curves, $K3$ surfaces and Fano 3–folds of genus ≤ 10*, from "Algebraic geometry and commutative algebra, I" (H Hijikata, H Hironaka, M Maruyama, H Matsumura, M Miyanishi, T Oda, K Ueno, editors), Kinokuniya, Tokyo (1988) 357–377 MR Zbl

[38] **S Mukai**, *Noncommutativizability of Brill–Noether theory and 3–dimensional Fano varieties*, Sūgaku 49 (1997) 1–24 MR Zbl

[39] **M Nagata**, *On self-intersection number of a section on a ruled surface*, Nagoya Math. J. 37 (1970) 191–196 MR Zbl

[40] **P E Newstead**, *Characteristic classes of stable bundles of rank 2 over an algebraic curve*, Trans. Amer. Math. Soc. 169 (1972) 337–345 MR Zbl

[41] **M Raynaud**, *Sections des fibrés vectoriels sur une courbe*, Bull. Soc. Math. France 110 (1982) 103–125 MR Zbl

[42] **X Sun**, *Moduli spaces of $\mathrm{SL}(r)$–bundles on singular irreducible curves*, Asian J. Math. 7 (2003) 609–625 MR Zbl

[43] **S-L Tan**, *On the slopes of the moduli spaces of curves*, Internat. J. Math. 9 (1998) 119–127 MR Zbl

[44] **M Teixidor i Bigas**, *Rank two vector bundles with canonical determinant*, Math. Nachr. 265 (2004) 100–106 MR Zbl

[45] **M Teixidor i Bigas**, *Petri map for rank two bundles with canonical determinant*, Compos. Math. 144 (2008) 705–720 MR Zbl

[46] **M Thaddeus**, *Conformal field theory and the cohomology of the moduli space of stable bundles*, J. Differential Geom. 35 (1992) 131–149 MR Zbl

[47] **A Verra**, *The unirationality of the moduli spaces of curves of genus 14 or lower*, Compos. Math. 141 (2005) 1425–1444 MR Zbl

[48] **C Voisin**, *Sur l'application de Wahl des courbes satisfaisant la condition de Brill–Noether–Petri*, Acta Math. 168 (1992) 249–272 MR Zbl

[49] **N Zhang**, *Towards the Bertram–Feinberg–Mukai conjecture*, J. Pure Appl. Algebra 220 (2016) 1588–1654 MR Zbl

Institut für Mathematik, Humboldt-Universität zu Berlin
Berlin, Germany

Department of Mathematics, University of Kentucky
Lexington, KY, United States

Department of Mathematics, University of Texas at Austin
Austin, TX, United States

farkas@math.hu-berlin.de,   dave.jensen@uky.edu,   sampayne@utexas.edu

Orbit equivalences of ℝ–covered Anosov flows and hyperbolic-like actions on the line

THOMAS BARTHELMÉ

KATHRYN MANN

APPENDIX WRITTEN JOINTLY WITH JONATHAN BOWDEN

We prove a rigidity result for group actions on the line whose elements have what we call "hyperbolic-like" dynamics. Using this, we give a rigidity theorem for ℝ–covered Anosov flows on 3–manifolds, characterizing orbit equivalent flows in terms of the elements of the fundamental group represented by periodic orbits. As consequences of this, we give an efficient criterion to determine the isotopy classes of self-orbit equivalences of ℝ–covered Anosov flows, and prove finiteness of contact Anosov flows on any given manifold.

In the appendix, with Jonathan Bowden, we prove that orbit equivalences of contact Anosov flows correspond exactly to isomorphisms of the associated contact structures. This gives a powerful tool to translate results on Anosov flows to contact geometry and vice versa. We illustrate its use by giving two new results in contact geometry: the existence of manifolds with arbitrarily many distinct Anosov contact structures, answering a question of Foulon, Hasselblatt and Vaugon, and a virtual description of the group of contact transformations of a Anosov contact structure, generalizing a result of Giroux and Massot.

37D20, 57M60

1 Introduction

1.1 Hyperbolic-like actions

A well-known theorem of Hölder states that any group acting freely by homeomorphisms of the line is abelian and conjugate to a group of translations. This was generalized in unpublished work of Solodov (see eg [Farb and Franks 2003; Kovačević 1999; Barbot 1995a]) to the statement that a group action on the line where each nontrivial element has at most *one* fixed point is either semiconjugate to an action by affine transformations, or abelian with a global fixed point. Later, the proof of the convergence group theorem [Gabai 1992; Casson and Jungreis 1994] established that a group action on the circle where each element has at most *two* fixed points is, under some additional technical dynamical hypotheses, conjugate to a subgroup of $\mathrm{PSL}(2, \mathbb{R})$ acting on $\mathbb{R}\mathrm{P}^1$ by Möbius transformations. This important result was the last step in the proof of the Seifert fiber space conjecture.

While one cannot reasonably expect further generalizations in this vein,[1] it is a natural question to ask what other fixed-point data might determine an action. In this spirit, we show that, under suitable hypotheses, an action of a group on the line is determined up to conjugacy by the set of elements acting with fixed points. Like the statement of the convergence group theorem, our hypotheses are motivated by an application to a classification problem, in our case the classification of $\mathbb{R}$–covered Anosov flows on 3–manifolds. Say that an action on the line is *hyperbolic-like* if it commutes with integer translation and each nontrivial element either acts freely or has exactly two fixed points in $[0, 1)$, one attracting and one repelling. We prove the following rigidity result for such actions:

Theorem 1.1 (rigidity of hyperbolic-like actions) *A minimal, hyperbolic-like action of a nonabelian group G on $\mathbb{R}$ is determined up to conjugacy by the set of elements of G that act with fixed points.*

Minimal and hyperbolic-like are both properties of the actions of 3–manifold fundamental groups on $\mathbb{R}$ induced by $\mathbb{R}$–covered Anosov flows. This allows us to use Theorem 1.1 to classify such flows up to orbit equivalence.

1.2 Orbit equivalence of Anosov flows

Recall that two flows on a manifold M are *orbit equivalent* if there is a homeomorphism $f : M \to M$ taking orbits of one to orbits of the other, and *isotopically equivalent* if this homeomorphism can be taken to be isotopic to the identity.

An Anosov flow is called $\mathbb{R}$–covered if the leaf space of its weak-stable foliation is homeomorphic to $\mathbb{R}$. (In the case of 3–manifolds, it is equivalent that the weak-unstable foliation has leaf space $\mathbb{R}$). On 3–manifolds, there are many constructions of $\mathbb{R}$–covered flows, and examples of manifolds admitting arbitrarily many inequivalent $\mathbb{R}$–covered flows. Here, we give a characterization of orbit and isotopy equivalent $\mathbb{R}$–covered flows on 3–manifolds by their *free homotopy classes of periodic orbits*. For a flow φ on a manifold M, let $\mathcal{P}(\varphi)$ denote the set of conjugacy classes of elements in $\pi_1(M)$ represented by the free homotopy classes of periodic orbits of φ. We show the following:

Theorem 1.2 (classification of $\mathbb{R}$–covered Anosov flows) *Let φ and ψ be $\mathbb{R}$–covered Anosov flows on a closed 3 manifold M.*

  (1) *φ and ψ are isotopically equivalent if and only if $\mathcal{P}(\varphi) = \mathcal{P}(\psi)$.*

  (2) *φ and ψ are orbit equivalent[2] if and only if there exists a homeomorphism $f : M \to M$ such that $f_*(\mathcal{P}(\varphi)) = \mathcal{P}(\psi)$. Moreover, the orbit equivalence can be taken to be in the isotopy class of f.*

---

[1] One reason for this is that the target groups $\mathbb{R}$, $\mathrm{Aff}_+(\mathbb{R})$ and $\mathrm{PSL}(2, \mathbb{R})$ are essentially the only Lie groups acting transitively on 1–manifolds, so the only natural candidates for such targets.

[2] In our definition of orbit equivalence, we do not require the homeomorphism to match *oriented* orbits to oriented orbits. If one wants to consider only orbit equivalences that preserve orbit orientation, then the conclusion of Theorem 1.2 will be that, if one flow is transversally orientable, then the other is also, and the orbit equivalence can be upgraded to an orientation-preserving orbit equivalence (using [Barbot 1995a], théorème C]). If one (and hence both) of the flows are not transversally orientable, then φ is orbit equivalent to either ψ or ψ^{-1}.

The ℝ–covered Anosov flows on closed 3–manifolds form a rich class of examples, including geodesic flows on closed surfaces, all contact Anosov flows, and diverse examples on many hyperbolic 3–manifolds and on manifolds with nontrivial JSJ decomposition (see [Fenley 1994; Foulon and Hasselblatt 2013; Barthelmé and Fenley 2017; Bonatti and Iakovoglou 2023; Bowden and Mann 2022]). Among other applications, our main theorem allows us to prove finiteness of contact Anosov flows. We describe and motivate the main applications now.

1.3 Application 1: classifying self-orbit equivalences

Describing the centralizer of a given diffeomorphism is a classical question in discrete-time dynamical systems; notably, Smale's conjecture [1998] is that the centralizer of a generic diffeomorphism should be trivial. By contrast, diffeomorphisms which embed in a flow have an ℝ–subgroup in their centralizer given by the flow, so the right analog of Smale's conjecture in this case is to ask whether the flow agrees (virtually) with the centralizer of the diffeomorphism. This motivates the general program to classify all symmetries of a given flow, and, more generally, classify the symmetries of the foliation given by orbits of a flow, ie classify the self-orbit equivalences. As with Smale's conjecture, this question is quite sensitive to regularity — for instance, 3–dimensional Anosov flows often have many self-orbit equivalences, while the set of those which may be realized by C^1 diffeomorphisms was shown to be virtually trivial by Barthelmé, Fenley and Potrie [Barthelmé et al. 2023].

Theorem 1.2 gives the following immediate characterization of isotopy classes of self-orbit equivalences:

Corollary 1.3 *A map* $f : M \to M$ *is in the isotopy class of a self-orbit equivalence of an* ℝ*–covered Anosov flow* φ *if and only if* $f_* : \pi_1(M) \to \pi_1(M)$ *preserves the set of conjugacy classes realized by periodic orbits of* φ.

With more work, we improve this to give a criterion to explicitly describe such classes, as follows. The case of interest here is for *skew* flows, those which are not orbit equivalent to a suspension of an Anosov diffeomorphism on the torus, as the orbit equivalences of suspension flows are essentially trivial.

If M is a 3–manifold with a skew Anosov flow φ, then M is orientable and irreducible, so admits a JSJ decomposition along tori into Seifert and atoroidal pieces. By Mostow rigidity and the structure of mapping class groups of Seifert spaces, the group Dehn(M) of isotopy classes of diffeomorphisms generated by Dehn twists along embedded tori in M has finite index in MCG(M) [Johannson 1979]. We give a criterion for when maps generated by certain Dehn twists represent a self-orbit equivalence of a flow. Given a flow φ and Dehn twist D_β, we may define an *orbit displacement* function, as follows. For each periodic orbit c in M, we set

$$d_\varphi(c, D_\beta) = \sum_{i=1}^{k} 2\epsilon_i t_\varphi(\beta_i),$$

where k is the number of transverse intersections of c with T (when T is in quasitransverse position with respect to the flow, $t_\varphi(\beta)$ is the *translation number* of the action of β on $\Lambda^s(\varphi)$, and $\epsilon_i = \pm 1$ is

an orientation term for each intersection. When f is a composition of Dehn twists D_{β_i} on disjoint nonisotopic tori, we define $d_\varphi(c, f)$ to be the sum of the displacements $d_\varphi(c, D_{\beta_i})$. Formal definitions are given in Section 3.

Theorem 1.4 (criterion for self-orbit equivalence) *Let φ be a transversally oriented skew Anosov flow. A map f which is a composition of Dehn twists on disjoint nonisotopic tori is isotopic to a self-orbit equivalence of φ if and only if, for all periodic orbits c of φ, we have $d_\varphi(c, f) = 0$.*

There are many situations in which this criterion is easy to check. As two sample applications, we have the following results for any transversally oriented skew Anosov flow φ:

Corollary 1.5 *Let D_{per} be the subgroup of the mapping class group of M generated by Dehn twists along curves represented by periodic orbits. Then any isotopy class $[h] \in D_{\mathrm{per}}$ is represented by a self-orbit equivalence of φ.*

Corollary 1.6 *Let T be an embedded torus and D_β a Dehn twist on T with nonzero translation number. Then D_β is isotopic to a self-orbit equivalence of φ if and only if T is separating in M.*

In Section 3.2, we discuss a number of other special cases where one may use the topology of M to reduce Theorem 1.4 to a simpler statement. We also discuss the complementary case to Theorem 1.4 for maps generated by Dehn twists in tori which cannot be realized disjointly, namely tori inside a single Seifert piece. See Theorems 3.17 and 3.18.

Remark 1.7 One can easily describe all self-orbit equivalences of a given skew Anosov flow in a fixed isotopy class. Such a flow comes with the data of a homeomorphism $\eta \colon M \to M$ realizing the *half-step-up* map on the orbit space. See Section 2.1 for details. It follows from [Barthelmé and Gogolev 2019, Theorem 1.1] that any two self-orbit equivalences h_1 and h_2 in the same isotopy class differ, up to isotopy along the flow lines, by some power of η.

Theorem 1.4 identifies which elements of a large subgroup of the mapping class group of M are represented by self-orbit equivalences. However, passing to the full mapping class group requires a different approach. In particular, we do not know the answer to the following:

Question 1.8 Does there exist an ($\mathbb{R}$–covered or not) Anosov flow on a hyperbolic 3–manifold M such that the only self-orbit equivalences are isotopic to the identity?

Question 1.9 Does there exist an ($\mathbb{R}$–covered or not) Anosov flow on a hyperbolic 3–manifold M such that every element of the mapping class group of M is represented by a self-orbit equivalence?

Remark 1.10 The case of most interest for both questions is when the manifold considered has nontrivial mapping class group. However, there are, as yet, no constructions of Anosov flows on a 3–manifold that has a trivial mapping class group. So even the trivial case for these questions is not yet known.

Remark 1.11   Barthelmé et al. [2023] introduced a class of partially hyperbolic diffeomorphisms, called "collapsed Anosov flows", which are semiconjugate to self-orbit equivalences of Anosov flows. Hence, the criterion of Theorem 1.4 (and its applications for certain manifolds, as in Theorems 3.17 and 3.18) describes the possible isotopy classes of collapsed Anosov flows associated with $\mathbb{R}$–covered Anosov flows.

1.4 Application 2: contact Anosov flows

An Anosov flow is said to be *contact* if it is the Reeb flow of a contact 1–form α. Notice that, with this definition, the contact structure $\xi = \ker \alpha$ is automatically transversely orientable since it is given as the kernel of a (globally defined) contact form. The contact Anosov flows are an important and well-studied class of examples, as they can be thought of as a generalization of the geodesic flow on manifolds of negative curvature, and many dynamical results on Anosov flows, for instance exponential decay of correlations [Liverani 2004], are known only for the contact case. In the context of 3–manifolds, Barbot [2001] proved that contact Anosov flows on 3–manifolds are necessarily $\mathbb{R}$–covered and skew, while Foulon–Hasselblatt surgery [Foulon et al. 2021] produces many examples. In fact, it is currently an open question whether every $\mathbb{R}$–covered skew flow is orbit equivalent to a contact flow. As progress towards a better understanding of these flows, we show that isomorphism of the associated contact structures is the same as orbit equivalence of flows, giving a powerful tool to use the machinery of flows to answer questions in contact geometry and vice versa.

Theorem 1.12   *Two contact Anosov flows on a 3–manifold are orbit equivalent if and only if their respective contact structures are contactomorphic. They are isotopically equivalent if and only if the contact structures are isotopic.*

Recall that two contact structures ξ_1 and ξ_2 on a manifold M are *contactomorphic* if there exists a diffeomorphism $g \colon M \to M$ such that $g_* \xi_1 = \xi_2$, and they are *isotopic* if g can be taken to be isotopic to the identity.

We prove the reverse of Theorem 1.12 in Section 4, and the forward direction in Theorem A.2. Using the coarse classification of tight contact structures of Colin, Giroux and Honda [Colin et al. 2009] and the result (proved in the appendix) that Anosov contact structures have zero torsion, we also obtain the following:

Theorem 1.13   (finiteness for contact Anosov flows)   *On any given 3–manifold M, there are only finitely many contact Anosov flows on M up to orbit equivalence.*

Thanks to Theorem 1.12, one can now fully translate results about contact Anosov flows to results about Anosov contact structures and vice versa. We illustrate this principle in the appendix with two

examples: First, we show (Theorem A.7) that there exists hyperbolic 3–manifolds with arbitrarily many noncontactomorphic Anosov contact structures, answering a question raised in [Foulon et al. 2021]. Second, we give a virtual description of the group of contact transformations of a Anosov contact structure up to isotopy on some manifolds (Theorem A.8). This generalizes a result by Giroux and Massot [2017].

Outline of the article

Section 2 gives a brief introduction to the structure of $\mathbb{R}$–covered flows on 3–manifolds, followed by the proof of Theorem 1.2. Section 3 contains the proof of Theorem 1.4, and the applications to contact flows are given in Section 4 and the appendix.

Acknowledgements

2 Proofs of Theorems 1.1 and 1.2

The first statement of Theorem 1.2 is a special case of the second, so we in fact only need to prove that assertion. Recall this is the statement that two flows φ and ψ are orbit equivalent if and only if there exists a homeomorphism $f : M \to M$ such that $f_*(\mathcal{P}(\varphi)) = \mathcal{P}(\psi)$, and, if this holds, the orbit equivalence can be taken to be in the isotopy class of f. The forward direction is immediate; we now set up the proof of the reverse direction. This will lead us to the statement and proof of Theorem 1.1; we finish the proof of Theorem 1.2 at the end of this section.

Throughout the work, we assume the reader has basic familiarity with Anosov flows. We recall below the essential structure theory of $\mathbb{R}$–covered flows on 3–manifolds that is used in the proof. Further background can be found in [Fisher and Hasselblatt 2019], and results specific to the topological theory of Anosov flows in dimension 3 can be found in [Barbot 2005].

By work of Fenley [1994] and Barbot [1995a], an $\mathbb{R}$–covered Anosov flow on a closed 3–manifold is either conjugate to the suspension of an Anosov diffeomorphism of T^2 or is *skew*, meaning that the orbit space of the lift of the flow to $\widetilde{M}$ is homeomorphic to the infinite diagonal strip

$$\mathbb{O} = \{(x, y) \in \mathbb{R}^2 : |x - y| < 1\}$$

via a homeomorphism taking the stable leaves of the flows to the horizontal cross-sections of the strip, and unstable leaves to the vertical cross-sections.
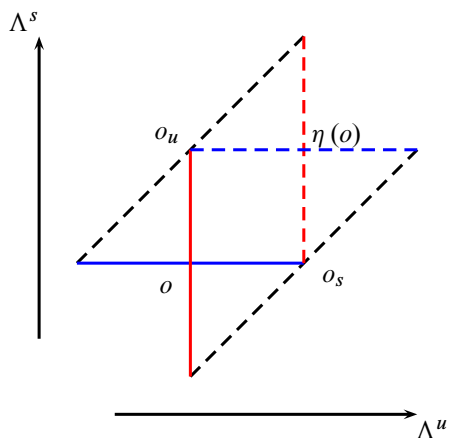
Figure 1: The orbit space $\mathbb{O}$.

[Theorem 1.2](#) follows from purely topological considerations in the suspension case, as follows. Suppose that M is a 3–manifold that fibers as the mapping torus of an Anosov diffeomorphism A on the torus. By a theorem of Plante [1981], any Anosov flow on M is necessarily of suspension type. Any other suspension flow comes from a fibering of M as the mapping torus of an Anosov diffeomorphism. It is a "folklore" result that such a diffeomorphism must be conjugate to either A or A^{-1} — a detailed proof can be found in [Funar 2013]. Since our definition of orbit equivalence allows a flow to be conjugate to its inverse, we conclude that M admits only one Anosov flow up to orbit equivalence.

The case of skew flows is much more interesting. For instance, examples of (closed, hyperbolic) manifolds that admit arbitrarily many inequivalent skew Anosov flows were constructed in [Bowden and Mann 2022]. As a first step towards the proof of [Theorem 1.2](#) in the skew case, we need to recall some general structure theory due to Barbot and Fenley that will allow us to essentially reduce the theorem to a statement about actions of $\pi_1(M)$ on S^1.

2.1 Skew Anosov flows

Consider again the infinite diagonal strip model for the orbit space as shown in [Figure 1](#).

In this model, each point $o \in \mathbb{O}$ can be assigned a point o_u on the upper boundary of the strip by following the unstable leaf through o, and a point o_l on the lower boundary by following the stable leaf. Taking the intersection of the stable leaf through o_u and unstable through o_l defines a continuous, fixed-point-free map $\eta \colon \mathbb{O} \to \mathbb{O}$, which we call the *half-step-up* map.[3] This map exchanges stable leaves and unstable leaves, so $\tau = \eta^2$ induces a map on the leaf space Λ^s of the weak stable foliation. We call τ the *one-step-up map*.

If the weak foliations are transversely orientable, then τ commutes with the action of $\pi_1(M)$ on Λ^s. Identifying $\Lambda^s \cong \mathbb{R}$ so that τ is identified with the translation $x \mapsto x + 1$ realizes $\pi_1(M)$ as a subgroup

---

[3]This map and its square have both been referred to as the one-step-up map elsewhere in the literature. We choose to call the square the one-step-up and η the half-step-up.

of $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$, the group of orientation-preserving homeomorphisms of $\mathbb{R}$ commuting with integer translations. In the nonorientable case, τ is twisted π_1–equivariant: for any $\gamma \in \pi_1(M)$ that reverses the orientation of Λ^s, we have $\tau(\gamma x) = \gamma \tau^{-1}(x)$.

Our perspective going forward will be to study the flows through the action of $\pi_1(M)$ on the leaf space. For simplicity, in Section 2.3 we assume first that the weak foliations are transversely orientable, and then state the necessary modifications for the nonorientable case. In the orientable case, the dynamics of the action of $\pi_1(M)$ on Λ^s is what Thurston [1997] calls an *extended convergence group action*. However, the only dynamical property that we will need is the fact that, in such a group, any orientation-preserving homeomorphism with fixed points has exactly two fixed points in $[0, 1)$, one attracting and one repelling. This can be seen directly from the Anosov dynamics of the flow.

Another dynamical property that will be of use comes from [Barbot 1995a, Theorem 2.5], which states that skew Anosov flows are transitive and the action of the group generated by $\pi_1(M)$ and τ on $\mathbb{R}$ is *minimal*, meaning that all orbits are dense. Since $\mathbb{R}/\tau \cong S^1$ and the action of $\pi_1(M)$ descends to this circle, this latter statement is equivalent to the statement that the action of $\pi_1(M)$ on $\mathbb{R}/\tau$ is minimal. The reader may find it useful to visualize the action on $\mathbb{R}$ by thinking of it on the circle $\mathbb{R}/\tau$, but for simplicity we will work in $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$. The next subsection establishes some general results about such subgroups of $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$, concluding with the proof of Theorem 1.1.

2.2 Hyperbolic-like homeomorphisms: proof of Theorem 1.1

Definition 2.1 We say an element $f \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ is *hyperbolic-like* if it has exactly two fixed points in $[0, 1)$, one attracting and one repelling, and a group action $\rho \colon G \to \mathrm{Homeo}(\mathbb{R})$ is hyperbolic-like if its image lies in $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ and every element with fixed points is hyperbolic-like.

In the context of the action of the fundamental group of a 3–manifold with a skew Anosov flow on the stable leaf space (which is what we have in mind), Thurston [1997] calls hyperbolic-like elements *space-like* homeomorphisms. Since this notation is not commonplace, we have chosen the terminology "hyperbolic-like" since the induced action of such elements on $\mathbb{R}/\mathbb{Z}$ are topologically conjugate to hyperbolic Möbius transformations.

If a is a hyperbolic-like element, we will use the notation a_+ and a_- to respectively denote an attracting and a repelling fixed point for a. For two hyperbolic-like elements a and b, we say the fixed sets of a and b are *linked* if each connected component of $\mathbb{R} \smallsetminus \mathrm{Fix}(a)$ contains a fixed point for b, or if a and b have fixed points in common. We say they are *unlinked* otherwise. By convention, when we speak about the order of fixed points on $\mathbb{R}$, we use the notation

$$a_+ < b_- < a_- < b_+$$

to mean that there exist four *consecutive* elements of $\mathrm{Fix}(a) \cup \mathrm{Fix}(b)$, ordered as indicated by the inequality. In this example, the fixed sets of a and b are linked. Since both a and b commute with integer translation,

the next element of $\mathrm{Fix}(a) \cup \mathrm{Fix}(b)$ to the right of b_+ is another attracting fixed point for a, equal to $a_+ + 1$.

The next series of lemmas shows that the configuration of fixed points of a pair or triple of elements can be detected by the set of words in those elements which act with fixed points.

Lemma 2.2 *Let* a *and* b *be hyperbolic-like elements of* $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$. *The fixed sets of* a *and* b *are linked if and only if* $\mathrm{Fix}(a^n b^m) \neq \varnothing$ *for all* $n, m \in \mathbb{Z}$.

Proof Suppose first that the fixed sets of a and b are linked, and let n and m be given. Since the property of having linked fixed sets does not change after passing to inverses, up to replacing a or b with their inverses we can assume that $n, m \geq 0$. Since the fixed sets of a and b are linked, either a and b have a common fixed point (in which case we are done) or there exists some connected component of $\mathbb{R} \smallsetminus (\mathrm{Fix}(a) \cup \mathrm{Fix}(b))$ bounded on one side by an attracting fixed point for a, and on the other by an attracting fixed point for b. Let I denote the closure of this component. Then $a^n b^m(I) \subset I$, so $a^n b^m$ has a fixed point in I.

To prove the converse, suppose now that the sets are unlinked. Up to passing to inverses, we may find consecutive attracting and repelling fixed points for a and b that lie in the order

$$a_+ < a_- < b_+ < b_-$$

with no other fixed points between a_+ and b_-. For m large enough, $b^m(a_+)$ will lie in the open interval (a_-, b_+). For n large enough, $a^n b^m(a_+)$ will therefore lie in the open interval $(b_-, a_+ + 1)$. Similarly, $b^m a^n b^m(a_+)$ will lie to the right of $a_+ + 1$, as will $a^n b^m a^n b^m(a_+)$. Thus, $(a^n b^m)^2$ translates some point a distance at least 1. Any such element of $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ is fixed-point-free; hence, its root $a^n b^m$ is fixed-point-free as well. $\square$

The next lemma says that, if a and b have unlinked fixed sets, then we can detect the cyclic order of attracting and repelling fixed points by understanding which words in a and b have fixed points.

Lemma 2.3 *Suppose* a *and* b *have unlinked fixed sets. The word* $b^N a^N$ *has a fixed point for every* $N > 0$ *if and only if one may find a set of consecutive fixed points either in the order*

$$a_+ < a_- < b_- < b_+$$

or that obtained from the above by replacing a *and* b *simultaneously with their inverses.*

Proof If the ordering shown above occurs, then the interval $[b_+, a_+ + 1]$ is mapped into itself by $b^N a^N$, so the map has a fixed point. Note that the ordering obtained by replacing a and b with their inverses can also be obtained simply by reversing the orientation of $\mathbb{R}$ and considering a sequence of consecutive fixed points starting with a_-. Since reversing orientation of $\mathbb{R}$ obviously does not change the property of

a map having fixed points, we have already proved one direction of the lemma. For the converse, if the ordering above does not occur even after passing to inverses, the order is either

$$a_+ < a_- < b_+ < b_- \quad \text{or} \quad a_- < a_+ < b_- < b_+.$$

In the first case, for sufficiently large N we have $(b^N a^N)^2(b_+) > b_+ + 1$, so the map is fixed-point-free, and in the second case we have $(b^N a^N)^2(b_+) < b_+ - 1$, so the map is again fixed-point-free. □

Our next goal is to use this information to reconstruct a minimal action, up to conjugacy, from the data of the ordering of fixed-point sets. For this we need an elementary lemma:

Lemma 2.4 *Suppose $G \subset \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ is a nonabelian group whose action on $\mathbb{R}$ is minimal and hyperbolic-like. Then, for any point $x \in \mathbb{R}$ and any $\epsilon > 0$, there exists $a \in G$ such that a has two fixed points in the ϵ–neighborhood of x.*

Proof If no nontrivial element of G acted with fixed points, then G would be abelian by Hölder's theorem, so by assumption this case does not occur. Thus, there exist hyperbolic-like elements and, by minimality of the action of G, the set of their attracting fixed points is dense.

Let x and $\epsilon > 0$ be given. Fix any hyperbolic-like element g with an attracting fixed point in the $\frac{\epsilon}{2}$–neighborhood of x. Observe that, if f and g are hyperbolic-like and f does not fix a repelling point g_- for g, then all fixed points of the conjugate $g^N f g^{-N}$ approach the attracting fixed points of g as $N \to \infty$. Thus, it suffices to find a hyperbolic-like f that does not fix g_-. By minimality, there exists $h \in G$ such that $h(g_-)$ lies strictly between g_- and g_+, where $g_- < g_+$ are consecutive fixed points. If $h(g_+ - 1) \neq g_-$, then hgh^{-1} is hyperbolic-like and has fixed points distinct from g_-, since they are the images of the fixed points of g under h, and we are done. If instead $h(g_+ - 1) = g_-$, or equivalently $h(g_+) = g_- + 1$, then we have

$$g_- < h^{-1}(g_+) < g_+ < g_- + 1$$

and therefore $h^{-1}gh$ has fixed points g_+ and $h^{-1}(g_+) \neq g_-$. □

Although not strictly needed in our proof, [Lemma 2.4](#) can be strengthened to the following density for pairs of fixed points:

Lemma 2.5 *Suppose $G \subset \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ is a nonabelian group whose action on $\mathbb{R}$ is minimal and hyperbolic-like. Given $x, y \in [0, 1)$, and $\epsilon > 0$, there exists a hyperbolic-like element $g \in G$ with fixed points satisfying $|g_- - x| < \epsilon$ and $|g_+ - y| < \epsilon$.*

Proof Let $x, y \in [0, 1)$ and $\epsilon > 0$ be given. Without loss of generality, assume $x < y$ and assume that ϵ is small enough that the ϵ–neighborhoods of x and y and all of their integer translates are pairwise disjoint.

By Lemma 2.4, we may find hyperbolic-like elements a and b with fixed points in the $\frac{\epsilon}{2}$–neighborhoods of x and y, respectively. Replacing a or b with their inverses if needed, we can assume these fixed points are ordered

$$a_+ < a_- < b_- < b_+.$$

By Lemma 2.3, this implies that $b^N a^N$ has a fixed point for every $N > 0$. Furthermore, if N is sufficiently large, an attracting fixed point for $b^N a^N$ will lie within the $\frac{\epsilon}{2}$–neighborhood of b_+, and a repelling fixed point within the $\frac{\epsilon}{2}$–neighborhood of a_-; this is simply because $b^N a^N$ takes a complement of the $\frac{\epsilon}{2}$–neighborhood of the union of translates of a_- to a neighborhood of the attracting fixed points for b. □

We can now finish the proof of our first main theorem; first we recall the statement:

Theorem 1.1 (rigidity of hyperbolic-like actions) *A minimal, hyperbolic-like action of a nonabelian group G on $\mathbb{R}$ is determined up to conjugacy by the set of elements of G that act with fixed points.*

Proof Let G be nonabelian, hyperbolic-like and acting minimally on $\mathbb{R}$, and $\rho(G)$ another such faithful action of G on $\mathbb{R}$, with the same set of hyperbolic-like elements as G. Let $g \in G$ be a hyperbolic-like element (recall that such an element exists by Hölder's theorem since G is nonabelian). Choose coordinates on $\mathbb{R}$ so that the attracting fixed points of g and of $\rho(g)$ are precisely the integers. We need to choose an orientation on the line $\mathbb{R}$ on which ρ acts. To do this, take some $f \in G$ with an axis unlinked with g. Such a map f exists by Lemma 2.4. Replacing f with its inverse if needed, we can find such a map for which consecutive fixed points are ordered

$$g_+ < g_- < f_- < f_+.$$

By Lemma 2.2 and the fact the action of $\rho(G)$ has the same elements with fixed points as the original action of G, we conclude that the fixed sets of $\rho(g)$ and $\rho(f)$ are unlinked. By the same reasoning, using Lemma 2.3, the map $f^N g^N$ has a fixed point for all $N > 0$, so by hypothesis the same is true for $\rho(f)^N \rho(g)^N$. Applying the other direction of Lemma 2.3, we can now fix an orientation on $\mathbb{R}$ so that consecutive fixed points of $\rho(f)$ and $\rho(g)$ are ordered

$$\rho(g)_+ < \rho(g)_- < \rho(f)_- < \rho(f)_+.$$

Our next goal is to show that this determines the ordering of the set of all attracting fixed points of hyperbolic-like elements of $\rho(G)$. We then define a map Θ on the (dense) subset of $\mathbb{R}$ consisting of attracting fixed points of other elements by sending the unique attracting fixed point of an element h that lies in $[m, m+1)$ to the unique attracting fixed point of $\rho(h)$ in $[m, m+1)$. This order-preserving property is sufficient to show that our map Θ is continuous, from which it will easily follow that it can be extended continuously to a homeomorphism of $\mathbb{R}$ that conjugates the action of G and $\rho(G)$.

Consider first an element h with fixed set that is unlinked with both f and g. Up to switching h with h^{-1}, there are three cases to consider.

Case 1 Suppose first that we have the ordering

$$g_+ < g_- < f_- < h_- < h_+ < f_+.$$

Applying Lemma 2.3, we have that $f^N h^N$ and $g^N h^N$ have fixed points for all positive N, and hence so do $\rho(f)^N \rho(h)^N$ and $\rho(g)^N \rho(h)^N$. Applying the lemma again implies that one of these three orderings occur:

(∗) $$\rho(g)_+ < \rho(h)_+ < \rho(h)_- < \rho(g)_- < \rho(f)_- < \rho(f)_+,$$

(∗∗) $$\rho(h)_+ < \rho(g)_+ < \rho(g)_- < \rho(h)_- < \rho(f)_- < \rho(f)_+,$$

(∗∗∗) $$\rho(g)_+ < \rho(g)_- < \rho(f)_- < \rho(h)_- < \rho(h)_+ < \rho(f)_+.$$

We want to show that only case (∗∗∗) can occur. We will show that (∗) does not occur. Eliminating the possibility of (∗∗) is done by exactly the same argument, switching the roles of g and h; we omit the details.

Since $\operatorname{Fix}(aba^{-1}) = a \operatorname{Fix}(b)$, for any $n > 0$ we also have

$$g_+ < g_- < f_- < (f^n h f^{-n})_- < (f^n h f^{-n})_+ < f_+$$

and so $g^N (f^n h f^{-n})^N$ has a fixed point for all $N > 0$, as does its image under ρ. If ordering (∗) were to occur, then for sufficiently large n we would have

$$\rho(g)_+ < \rho(g)_- < \rho(f)_- < \rho(f)_+ < \rho(f^n h f^{-n})_+ < \rho(f^n h f^{-n})_-,$$

contradicting Lemma 2.3 applied to $\rho(g)^N \rho(f^n h f^{-n})^N$. Thus, the ordering of consecutive fixed points of f, g and h under ρ agrees with that for the original action.

Case 2 The ordering

$$g_+ < h_- < h_+ < g_- < f_- < f_+.$$

is handled exactly as above, exchanging the roles of f and g.

Case 3 Now suppose instead that the ordering of the fixed points of f, g and h is

$$g_+ < g_- < h_+ < h_- < f_- < f_+.$$

Consider the elements $a = g^n f^n$ and $b = f^n h^n$ for some large positive n. As $n \to \infty$, the attracting fixed point a_+ approaches f_+, and similarly a_- approaches g_-, b_+ approaches h_+, and b_- approaches f_-. Thus, provided n is chosen large enough, we have

$$g_+ < g_- < a_- < b_+ < h_+ < h_- < b_- < f_- < f_+ < a_+.$$

We can then apply the previous cases to the triples (g, a, h), (g, a, f), (g, a, b) and (g, b, h) to show the ordering of their fixed points is preserved by ρ. We deduce that the ordering of the fixed points of $\rho(f)$, $\rho(g)$ and $\rho(h)$ matches that of the fixed points of f, g, and h.

We can now quickly finish the proof of the theorem. Suppose we have some hyperbolic-like a and b with

$$0 \leq a_+ < b_+ < 1.$$

Rather than consider cases depending on whether a and b are linked or not, we can instead use Lemma 2.4 to choose c and d in G with fixed sets very close to a_+ and b_+ such that c and d have fixed sets unlinked with a and b. Suppose for simplicity that $0 < a_+$ (otherwise, simply replace 0 in what follows with some very small $-\epsilon$ and 1 with $1 - \epsilon$, and repeat the proof). By Lemma 2.5, we can choose elements c and d with fixed points so that we have the ordering

$$0 < c_- < c_+ < a_+ < d_- < d_+ < b_+ < 1$$

and additionally have that a does not have a repelling fixed point between c_- and a_+, and b does not have a repelling fixed point between d_- and d_+. We can also choose such c and d whose fixed sets are each unlinked with respect to f and to g, so that we may apply the observation above to c and d and determine the relative order of their fixed-point sets. Any choice of c and d with fixed sets sufficiently close to a_+ and b_+ will have this property. Thus, by our convention on $\rho(g)$, we conclude that

$$\rho(g)_+ = 0 < \rho(c)_- < \rho(c)_+ < \rho(d)_- < \rho(d)_+ < 1.$$

Since c and d had unlinked fixed points with respect to a, b and g, we can apply the observation again and conclude that the ordering of fixed sets is preserved, namely

$$\rho(g)_+ = 0 < \rho(c)_- < \rho(c)_+ < \rho(a)_+ < \rho(d)_- < \rho(d)_+ < \rho(b)_+ < 1$$

and, in particular, $\rho(g)_+ = 0 < \rho(a)_+ < \rho(b)_+ < 1$, as we needed to show.

Thus, we have defined an order-preserving (and hence continuous) injective map between two dense subsets of $\mathbb{R}$. This extends uniquely to a continuous map $\mathbb{R} \to \mathbb{R}$ with continuous inverse, which we denote by Θ. It remains to see that Θ conjugates the actions of G and $\rho(G)$. Let $g \in G$ be given. Note that, if a_+ is a fixed point of a hyperbolic-like element of $a \in G$, then $g(a_+)$ is an attracting fixed point of (gag^{-1}); thus, $\Theta(ga_+)$ is some attracting fixed point of $\rho(gag^{-1})$, ie the image of an attracting fixed point of $\rho(a)$ under $\rho(g)$. In other words, $\Theta(ga_+) = \rho(g)\Theta(a_+) + n$ for some $n \in \mathbb{Z}$. Since Θ is continuous and fixed points of hyperbolic-like elements are dense in the source and the target, we conclude that $n = 0$, so g–equivariance holds on a dense set, and hence everywhere. $\qquad \square$

Remark 2.6 If G is not assumed to act minimally (but $\rho(G)$ is), the same proof strategy can be used to produce a *semiconjugacy* between the actions of G and $\rho(G)$, defined on the closure of the G–invariant set consisting of hyperbolic fixed points.

2.3 Conclusion of the proof of Theorem 1.2

Returning to the setup of Theorem 1.2, suppose that φ and ψ are two skew Anosov flows with $f_*\mathcal{P}(\varphi) = \mathcal{P}(\psi)$ for some homeomorphism $f \colon M \to M$. Replacing φ with its conjugate under f, we obtain a

flow φ satisfying $\mathcal{P}(\varphi) = \mathcal{P}(\psi)$. What we need to show is that φ and ψ are orbit equivalent by some homeomorphism of M that is isotopic to the identity.

Orientable case Assume first that $\Lambda^s(\varphi)$ and $\Lambda^s(\psi)$ are both transversely orientable. We then will apply Theorem 1.1 to the group $G := \pi_1(M)$. Recall the action of this group on the leaf space $\Lambda^s(\varphi) \cong \mathbb{R}$ is faithful, minimal and has the property that all elements with fixed points are hyperbolic-like, under the parametrization of $\mathbb{R}$ where $\tau(\varphi)$ acts as translation by 1, and the same holds for ψ.

That $\mathcal{P}(\varphi) = \mathcal{P}(\psi)$ means precisely that these two representations have the same elements with fixed points. Thus, Theorem 1.1 implies that these two actions are conjugate. This also gives a conjugacy between the actions on the unstable leaf spaces $\Lambda^u(\varphi)$ and $\Lambda^u(\psi)$ via further conjugation by the half-step-up map η. Considering intersections of stable and unstable leaves, we can promote this to a conjugacy of the actions of $\pi_1(M)$ on the orbit space $\mathcal{O}$. Following an argument of Ghys using Haefliger's theory of classifying spaces of foliations, Barbot [1995a, Theorem 3.4] showed using an averaging trick that such a conjugacy on the orbit space can always be realized by a homeomorphism of M giving an orbit equivalence of the flows. In our case, this homeomorphism is easily seen to be isotopic to the identity by considering the action on π_1. This concludes the proof in the transversely orientable case.

General case For the general case, consider again the action of $\pi_1(M)$ on $\Lambda^s(\varphi) \cong \mathbb{R}$ and on $\Lambda^s(\psi)$. Let $G \subset \pi_1(M)$ be the normal subgroup generated by all squares of elements. Since each element of G is a product of squares, its action on $\Lambda^s(\varphi)$ and $\Lambda^s(\psi)$ is by orientation-preserving homeomorphisms. The proof of [Barbot 1995a, théorème 2.5] shows directly that the action of G on $\Lambda^s(\varphi)$ and $\Lambda^s(\psi)$ is also minimal. Thus, we may apply Theorem 1.1 and conclude that the actions of G on the respective leaf spaces are conjugate. We wish to show that this conjugacy extends to a conjugacy of the actions of $\pi_1(M)$.

Apply a conjugacy so that the actions of G on $\Lambda^s(\varphi)$ and $\Lambda^s(\psi)$ agree. Now our goal is to show that, after possibly further conjugating by an integer translation (which commutes with the action of G), the actions of $\pi_1(M)$ agree. Let ρ_φ and ρ_ψ denote the actions on $\Lambda^s(\varphi)$ and $\Lambda^s(\psi)$, respectively, assumed to agree on the restrictions to G.

Note that, if $\gamma \in \pi_1(M)$, and $x \in \mathbb{R}$ is an attracting fixed point of $\rho_\varphi(g)$ for some element $g \in G$, then $\rho_\varphi(\gamma)(x)$ is an attracting fixed point of $\rho_\varphi(\gamma g \gamma^{-1})$, where we have $\gamma g \gamma^{-1} \in G$. The same applies to $\rho_\psi(\gamma)$. The set of all attracting fixed points for elements of G is dense, and each element with fixed points has a $\mathbb{Z}$–invariant set of attracting fixed points with exactly one in $[0, 1)$.

First we verify that $\rho_\varphi(\gamma)$ preserves orientation if and only if $\rho_\psi(\gamma)$ does. Suppose $\rho_\varphi(\gamma)$ reverses orientation. Let g_1, g_2 and g_3 in G be elements with attracting fixed points satisfying

$$\rho_\varphi(g_1)_+ < \rho_\varphi(g_2)_+ < \rho_\varphi(g_3)_+ < \rho_\varphi(g_1)_+ + 1.$$

Since $\rho_\varphi(\gamma)$ reverses orientation, we have

$$\rho_\varphi(\gamma g_1 \gamma^{-1})_+ > \rho_\varphi(\gamma g_2 \gamma^{-1})_+ > \rho_\varphi(\gamma g_3 \gamma^{-1})_+ > \rho_\varphi(\gamma g_1 \gamma^{-1})_+ - 1$$

whereas, if $\rho_\varphi(\gamma)$ preserved the orientation, the original order would not be affected by conjugation. Since $\gamma g \gamma^{-1} \in G$, this ordering holds also for the action under ρ_ψ, showing that $\rho_\psi(\gamma)$ necessarily reverses orientation as well. The situation being symmetric, we have proved the claimed "if and only if" statement.

Now consider the subgroup P of elements of $\pi_1(M)$ whose action preserves orientation on $\Lambda^s(\varphi)$, or, as we have just shown, equivalently preserves orientation of $\Lambda^s(\psi)$. Our description above also implies that, for any $\gamma \in P$, we have $\rho_\varphi(\gamma) = \rho_\psi(\gamma) \circ T_\gamma$ for some integer translation T_γ. Since orientation-preserving elements commute with integer translation, the map $\gamma \in P \mapsto T_\gamma$ is a group homomorphism. Now, when $\gamma \in G$, T_γ is the identity; thus, T_γ is the identity for any $\gamma \in P$. So, in particular, the actions of ρ_φ and ρ_ψ are identical on P.

If instead we consider γ an element reversing the orientation on the leaf spaces, then $\rho_\psi(\gamma)$ and $\rho_\varphi(\gamma)$ each have a unique fixed point. Their action being determined modulo integer translations means that there exists an integer translation T'_γ such that $\rho_\varphi(\gamma) = T'_\gamma \rho_\psi(\gamma) T'^{-1}_\gamma$. Fix some orientation-reversing element γ_0. Up to conjugating the action of ρ_ψ by an integer translation, we can assume that T'_{γ_0} is the identity, ie that $\rho_\psi(\gamma_0) = \rho_\varphi(\gamma_0)$. Notice that this conjugation does not affect the fact that ρ_φ and ρ_ψ are identical on P, since the action of elements in P commutes with integer translations. We wish to show now that $T'_\gamma = 0$ for all $\gamma \in \pi_1(M) \smallsetminus P$, so the actions agree.

Let γ be any orientation-reversing element. Then $\gamma_0\gamma$ preserves the orientation so $\rho_\psi(\gamma_0\gamma) = \rho_\varphi(\gamma_0\gamma)$. As $\rho_\psi(\gamma_0) = \rho_\varphi(\gamma_0)$, we deduce directly that $\rho_\psi(\gamma) = \rho_\varphi(\gamma)$. Hence, we proved that the actions ρ_φ and ρ_ψ are the same, ending the proof of Theorem 1.2.

3 Classifying self-orbit equivalences

This section gives the proof of Theorem 1.4. We start by introducing some additional necessary background material on the orbit space of the flow.

Returning to the picture from Section 2.1, recall that the orbit space of a skew Anosov flow is homeomorphic to a diagonal strip in $\mathbb{R}^2$ foliated by Λ^s and Λ^u in the two coordinate directions. For each orbit $o \in \mathbb{O}$, the ideal quadrilateral in $\mathbb{O}$ with corners o and $\eta(o)$ and sides the stable and unstable (half-)leaves of o and $\eta(o)$ is called a *lozenge*. The union of lozenges associated with the orbits $\eta^n(o)$ for $n \in \mathbb{Z}$ is called a *string of lozenges*. The reader may consult [Barthelmé and Fenley 2017, Section 2] for more background about lozenges in general Anosov flows.

Recall [Barbot 1995b] that an (immersed) incompressible torus T is *quasitransverse* to an Anosov flow φ if it is transverse everywhere except along finitely many periodic orbits of φ. Barbot [1995b, théorème C] (see also [Barbot and Fenley 2013, Theorem 6.10]) showed that any incompressible embedded torus in M can be isotoped to a quasitransverse torus unless it is the boundary of a tubular neighborhood of an embedded one-sided Klein bottle. Such embeddings of a Klein bottle do not actually arise if the flow is
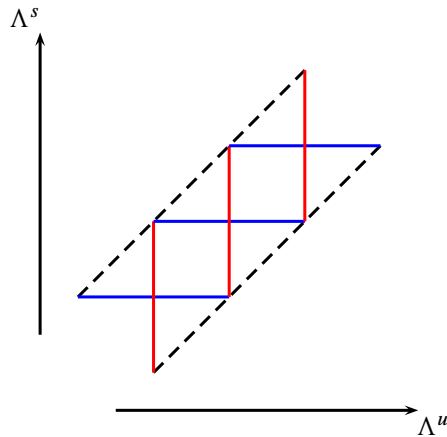
Figure 2: The orbit space $\mathcal{O}$ with (part of) a string of lozenges.

transversally orientable and $\mathbb{R}$–covered. We were not able to locate a proof of this fact in the literature, so add it as a lemma below. We thank Sergio Fenley for discussing this fact with us.

Lemma 3.1 *Let φ be a transversally orientable skew $\mathbb{R}$–covered Anosov flow on M. Then M does not contain a π_1–injective immersion of the Klein bottle.*

Proof Suppose there is a subgroup G of $\pi_1(M)$ isomorphic to the fundamental group of the Klein bottle, ie we have nontrivial elements $a, b \in \pi_1(M)$ such that $aba^{-1} = b^{-1}$ and $G = \langle a, b \rangle$. Since φ is transversely orientable, G acts faithfully on $\Lambda^s(\varphi) \simeq \mathbb{R}$ by orientation-preserving homeomorphisms commuting with integer translations. Consider the induced action of G on $\mathbb{R}/\eta^2$. For this induced action, both a and b have a single attracting and single repelling fixed point. Since $aba^{-1} = b^{-1}$, we must have that a interchanges the attracting and repelling points for b. But this gives a an orbit of period two, contradicting that it acts with fixed points. $\square$

The main result that we need is the following:

Lemma 3.2 [Barbot 1995b] *If T is a quasitransverse torus in M for a skew Anosov flow, then each lift $\widetilde{T}$ of T to the universal cover projects to a unique string of lozenges $\mathscr{C}$ in $\mathcal{O}$. That string of lozenges is the unique string of lozenges left invariant by the $\mathbb{Z}^2$–subgroup of $\pi_1(M)$ that fixes $\widetilde{T}$. The interior of each lozenge in the string corresponds to a maximal annulus in T transverse to the flow; its corners are the periodic orbits bounding the annulus.*

Conversely, if G is a subgroup of $\pi_1(M)$ isomorphic to $\mathbb{Z}^2$, then there exists a unique string of lozenges in $\mathcal{O}$ fixed by G. Moreover that string of lozenge is the projection of an (a priori only immersed) torus T.

Remark 3.3 We have stated this lemma in the special case of skew Anosov flows. For general Anosov flows on 3–manifolds, one needs to replace "string of lozenges" by "chain of lozenges" — see [Barthelmé and Fenley 2017, Section 2] — and some $\mathbb{Z}^2$–subgroups may fix two distinct chains of lozenges.

We also need the following easy proposition, giving another condition for an element $\gamma \in \pi_1(M)$ to be represented by a periodic orbit of a skew Anosov flow φ. Recall that $\Lambda^s \cong \mathbb{R}$ denotes the leaf space of the weak stable foliation of the lifted flow $\widetilde{\varphi}$. We denote the natural projection by $\pi^s \colon \widetilde{M} \to \Lambda^s$.

Proposition 3.4 *Let φ be a skew Anosov flow on a 3–manifold M. Let $\gamma \in \pi_1(M)$. The following are equivalent*:

(i) *The element γ is represented by a periodic orbit of φ;*

(ii) *There exists a γ–invariant curve $c \subset \widetilde{M}$ such that $\pi^s(c)$ is bounded (either above or below) in Λ^s;*

(iii) *For any γ–invariant curve $c \subset \widetilde{M}$, its image $\pi^s(c)$ is bounded in Λ^s.*

Proof Item (iii) trivially implies (ii), we will show that (ii) implies (i) and that (i) implies (iii).

(ii) $\Longrightarrow$ (i) Assume that there exists a γ–invariant curve $c \subset \widetilde{M}$ such that $\pi^s(c)$ is bounded above in Λ^s, and let L^+ denote its upper bound. (The case where $\pi^s(c)$ is bounded below is analogous.) If the action of γ reverses the orientation of Λ^s, then it fixes a leaf which necessarily contains a periodic orbit represented by γ, and we are done. If γ instead preserves the orientation, then the upper bound $L^+ \in \Lambda^s$ is a γ–invariant point since $\pi^s(c)$ is γ–invariant, and thus the leaf corresponding to this point contains a unique γ–invariant orbit. This shows (i).

(i) $\Longrightarrow$ (iii) Suppose γ is represented by a periodic orbit of the flow and $c \subset \widetilde{M}$ is a γ–invariant curve. Supposing first that γ preserves orientation on $\mathbb{R}$, the action has an attracting fixed point $x \in \Lambda^s$. Moreover, the images of this point under powers of the half-step-up map are alternately attracting and repelling fixed points of γ; if i is even, then $\eta^i(x)$ is an attracting fixed point, if i is odd, then $\eta^i(x)$ is a repelling fixed point, as described in Section 2.1. In the case where γ reverses orientation, the same argument above applies if we replace γ with γ^2, which preserves orientation.

In either case, since the action of γ^2 on c is free, we may take a compact fundamental domain I for the action. The projection $\pi^s(I)$ is contained in some bounded interval $[\eta^{-k}(x), \eta^k(x)]$. Therefore, $\pi^s(c) = \bigcup_{n \in \mathbb{Z}} \gamma^{2n} \cdot \pi^s(I) \subset [\eta^{-k}(x), \eta^k(x)]$, so it is bounded, proving (iii).　　　□

Combining Theorem 1.2 with Proposition 3.4 gives the following as a direct consequence:

Theorem 3.5 *Let φ and ψ be two skew $\mathbb{R}$–covered Anosov flows on a 3–manifold M. The flows φ and ψ are isotopically equivalent if and only if, for any periodic orbit α of ψ (resp. φ) with lift $\widetilde{\alpha} \subset \widetilde{M}$, the projection $\pi_1^s(\widetilde{\alpha}) \subset \Lambda^s(\varphi)$ (resp. $\pi_1^s(\widetilde{\alpha}) \subset \Lambda^s(\psi)$) is bounded.*

3.1 Proof of Theorem 1.4

To set up for the proof, we begin by giving precise definitions of translation number with respect to φ, and the displacement of a curve by a Dehn twist.

Definition 3.6 (translation number)   Let $\beta \in \pi_1(M)$, and consider its action on $\Lambda^s(\varphi)$. Fix $x \in \Lambda^s(\varphi)$. For each $q \in \mathbb{Z}$, there exists a unique $p_q \in \mathbb{Z}$ such that $\tau^{p_q}(x) \leq \beta^q(x) < \tau^{p_q+1}(x)$. We define

$$t_\varphi(\beta) := \lim_{q \to \infty} \frac{p_q}{q}.$$

Since τ and β commute, it is a standard exercise to show that this limit exists and is in fact independent of the choice of x; indeed, $t_\varphi(\beta)$ is simply the classical *translation number* for the action of β on Λ_s with respect to any parametrization of Λ_s where τ acts as translation by 1.

Remark 3.7   We will consider below only the case when β is freely homotopic to a curve in a quasi-transverse torus T. Thus, β fixes a (unique) string of lozenges in the orbit space. In this case we therefore have $t_\varphi(\beta) = k \in \mathbb{Z}$, where k is such that, if x is any of the corner of the string of lozenges, then $\beta x = \tau^k(x)$.

If a skew Anosov flow φ on M is transversally oriented, we saw (Lemma 3.1) that M cannot admit an incompressible embedding of the Klein bottle; thus, by [Barbot 1995b, théorème C], any incompressible embedded torus in M can be put in quasitransverse position with respect to the flow. We further have, thanks to [Barbot 1995b, théorème E], that any collection of pairwise disjoint, nonisotopic incompressible embedded tori can be simultaneously isotoped to a collection of still disjoint (and obviously still nonisotopic) quasitransverse tori. It is thus no loss of generality to adopt the following convention for transversely orientable flows:

Convention   For the remainder of this section, we restrict our attention to transversely orientable flows, and we will always assume that the tori we consider are in quasitransverse position.

Next we will define the "displacement" of an orbit by a Dehn twist. We first recall the definition of Dehn twists to emphasize that they come with a specification of a transverse orientation on the torus.

Definition 3.8   Let T be an embedded torus, and $\beta \in \pi_1(T)$. A *Dehn twist along* β is the mapping class of a map D_β defined as follows: Take a small product region $T \times [-1, 1] \in M$, and fix a basis $\{\alpha, v\}$ for $\pi_1(T)$ giving an identification of T with $S^1 \times S^1 = \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$, where $\beta = \alpha^p v^q$ for some p and q. For $(x, y, z) \in T \times [-1, 1]$, we define $D_\beta(x, y, t) = (x + ph(t), y + qh(t), t)$, where $h: [-1, 1] \to [0, 1]$ is a smooth bump function with $h(-1) = 0$ and $h(1) = 1$, and extend D_β to be the identity elsewhere on M.

Remark 3.9   Reversing the transverse orientation of T and applying the same construction results in a map isotopic to the *inverse* of that defined above. Thus, the notation D_β, while standard, is somewhat misleading because the mapping class does not depend on β alone. There is no intrinsic way to distinguish D_β from D_β^{-1}. Thus, by convention, we say that a map D_β *comes with* the data of a choice of transverse orientation. When we speak of a Dehn twist supported on a torus neighborhood $T \times [-1, 1]$, we always assume the orientation is as given by the interval $[-1, 1]$.

It will be convenient for us to choose homeomorphisms representing Dehn twists which are in a particularly nice form with respect to the flow, as follows:

Convention 3.10 (good Dehn twist coordinates) Given a quasitransverse torus T, we choose the coordinates $(x, y, z) \in T \times [-1, 1]$ in the following way: Let $\alpha_1, \ldots, \alpha_{2n}$ be the periodic orbits of φ on T. Then we assume that the local stable leaves of the orbits α_i in $T \times [-1, 1]$ are given by the equation $x = i/2n$.

With this convention, a Dehn twist D_α on T, where $\alpha \in \pi_1(T)$ represents any power of the periodic orbits, will preserve the local stable leaves of the periodic orbits α_i. More generally, given any Dehn twist D_β on T, and any segment $c(t) := (x_0, y_0, t)$ for $t \in [-1, 1]$ through $T \times [-1, 1]$, the number of times its image $D_\beta(c(t))$ intersects the union of stable leaves of the α_i is the minimal intersection number of β with α_i.

Definition 3.11 (sign of an intersection) Given a Dehn twist D_β supported on $T \times [-1, 1]$ as above, and an orbit φ_t intersecting T transversely at $t = 0$, we say this intersection is *positive* if $\varphi^t(z) \in T \times [0, 1]$ for small positive t, and *negative* if $\varphi^t(z) \in T \times [-1, 0]$ for small positive t.

Before giving the formal definition of the displacement $d_\varphi(c, f)$ of a closed orbit c under a product of Dehn twists f, we motivate this with the following lemma, which describes how a Dehn twist on a torus T affects a segment of a periodic orbit transverse to T, from the perspective of the leaf space of φ.

For the statement, we fix a quasitransverse torus T in M, a point z with orbit $\varphi^t(z)$ transverse to T, and a small product neighborhood $T \times [-1, 1]$ so that the orbit $\varphi^t(z) \cap T \times [-1, 1]$ is, locally near $t = 0$, a segment J between some point $z_- \in T \times \{-1\}$ and $z_+ \in T \times \{1\}$. Let D_β be a Dehn twist supported on $T \times [-1, 1]$.

Lemma 3.12 *Let $\widetilde{T} \times [-1, 1]$ be a lift of $T \times [-1, 1]$ to $\widetilde{M}$, let $\widetilde{J}$ be the lift of J in $\widetilde{T} \times [-1, 1]$ with endpoints $\tilde{z}_-$ and $\tilde{z}_+$, and let $\widetilde{D}_\beta$ be the lift of D_β fixing $\tilde{z}_-$. Finally, let $\{L_i\}_{i \in \mathbb{Z}}$ be the string of lozenges associated with $\widetilde{T}$.*

(1) *If $\widetilde{J}$ projects to a point in the lozenge L_i, then $\widetilde{D}_\beta(\tilde{z}_+)$ projects into L_k, where $k = i + 2t_\varphi(\beta)$ if the intersection of J and T is positive, and $k = i - 2t_\varphi(\beta)$ if the intersection is negative.*

(2) *The stable saturation of $\widetilde{D}_\beta(\widetilde{J})$ to the orbit space stays inside the stable saturation of the lozenges between L_i and L_k.*

Proof Without loss of generality, we assume that $\widetilde{J}$ projects to a point in L_0, and that the sign of the intersection is positive (the negative case is analogous). Recall that the lozenges L_i are the projections of strips A_i inside $\widetilde{T}$, bounded by periodic orbits of $\widetilde{\varphi}$.

By definition of D_β, the image $\widetilde{D}_\beta(\tilde{z}_+)$ projects in the orbit space to the lozenge $\beta(L_0)$. Now, if x_i are the corners of the lozenge L_i (enumerated so that $x_i < x_{i+1}$ in $\Lambda_s(\varphi)$), then $\eta^k(x_0) = x_{2k}$. Thus, by definition of translation number (and Remark 3.7), $\beta(L_0) = L_{2t_\varphi(\beta)}$, proving the first part of the lemma.

The second statement follows immediately from Convention 3.10 and the structure of lozenges (Lemma 3.2). $\qquad\square$
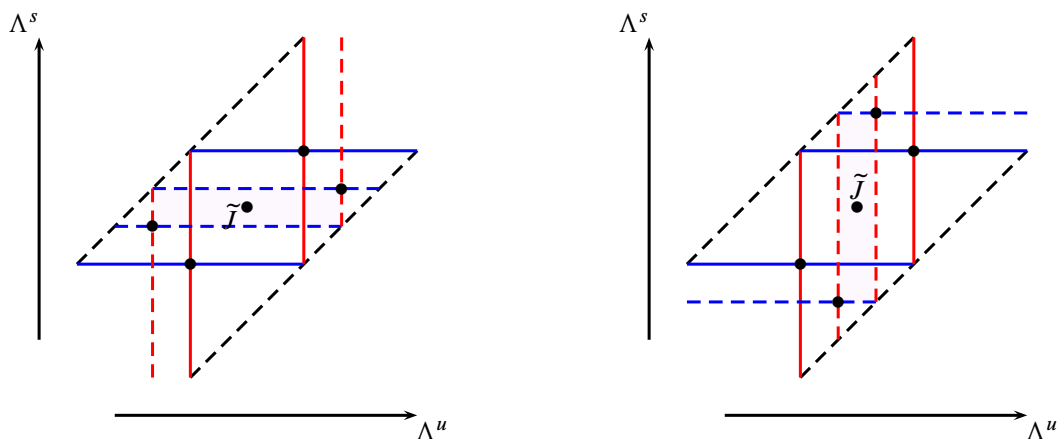
Figure 3: The configuration on the right is impossible if the shaded lozenge is L'_j.

Definition 3.13 (displacement)   Let c be a periodic orbit of the skew flow φ and D_β a Dehn twist on a quasitransverse torus T. Let $x_1, \ldots, x_n$ be the intersection points of c with T, and let $\epsilon_i = \{\pm 1\}$ be the sign of the intersection at x_i. The *displacement of c by D_β* is defined by

$$d_\varphi(c, D_\beta) = \sum_{i=1}^{n} 2\epsilon_i t_\varphi(\beta).$$

If $f = D_{\beta_1} \circ \cdots \circ D_{\beta_k}$ is a composition of Dehn twists on *pairwise disjoint* quasitransverse tori, then we set

$$d_\varphi(c, f) = \sum_{i=1}^{k} d_\varphi(c, D_{\beta_i}).$$

We will need the following observation for the proof of Theorem 1.4:

Observation 3.14   *Let $\tilde{T}$ and $\tilde{T}'$ be two **disjoint** lifts of* (*the same or distinct*) *quasitransverse tori. Let $\tilde{J}$ be a segment of a periodic orbit that first crosses $\tilde{T}$ and then crosses $\tilde{T}'$. Suppose $\tilde{T}$ projects to a string of lozenges $\{L_i\}$, and $\tilde{T}'$ to a string $\{L'_i\}$. Let i and j be such that $\tilde{J}$ is contained in $L_i \cap L'_j$.*

Then L'_j is contained in the saturation of L_i by stable leaves; equivalently, L_i is contained in the saturation of L'_j by unstable leaves.

Proof   Since the lifts $\tilde{T}$ and $\tilde{T}'$ are disjoint, observe that either $L_i \subset \widetilde{\mathcal{F}}^s(L'_j)$ and $L'_j \subset \widetilde{\mathcal{F}}^u(L_i)$, or $L_i \subset \widetilde{\mathcal{F}}^u(L'_j)$ and $L'_j \subset \widetilde{\mathcal{F}}^s(L_i)$. Both configurations are shown in Figure 3.

We will show that the second case cannot occur because orbits converge in the positive direction of the flow along $\tilde{J}$. Precisely, consider the strong unstable leaf through the starting point $\tilde{x}$ of the orbit segment $\tilde{J}$. Denote this leaf by λ^{uu}, and let $l_i \subset \lambda^{uu}$ denote the set of points whose forward orbit intersects the strip A_i in $\tilde{T}$ corresponding to L_i. Similarly, define $l'_j \subset \lambda^{uu}$ to be the set of points whose forward

orbit intersects the strip A_j' in $\widetilde{T}'$ corresponding to L_j'. Hyperbolicity of the flow means that $l_j' \subset l_i$. Hence, $L_j' \subset \widetilde{\mathcal{F}}^s(L_i)$, and equality happens if and only if $L_j' = L_i$, which in turn implies that the orbits bounding A_i and A_j' are the same, which contradicts the assumption that $\widetilde{T}$ and $\widetilde{T}'$ are quasitransverse and disjoint. $\qquad\square$

We next reduce the proof of Theorem 1.4 to the following proposition:

Proposition 3.15 *Suppose $f = D_{\beta_1} \circ \cdots \circ D_{\beta_n}$ is a composition of Dehn twists on pairwise disjoint quasitransverse tori $T_1, \ldots, T_n$, and let $\tilde{c} \subset \widetilde{M}$ be a lift of a periodic orbit c of φ. Then $d_\varphi(c, f) = 0$ if and only if, for some lift (equivalently, for all lifts) $\tilde{f}$ of f, the image $\tilde{f}(\tilde{c})$ has bounded projection to $\Lambda^s(\varphi)$.*

Proof of Theorem 1.4 given Proposition 3.15   Indeed, this is an easy consequence of Proposition 3.4 and Theorem 1.2. Theorem 1.2 states that f is a self-orbit equivalence if and only if $\mathcal{P}(\varphi) = f_*\mathcal{P}(\varphi) = \mathcal{P}(f\varphi f^{-1})$. By Proposition 3.4, $\mathcal{P}(\varphi)$ can be characterized as the set of $\gamma \in \pi_1(M)$ such that every γ–invariant curve $\tilde{c} \subset \widetilde{M}$ has bounded projection to $\Lambda^s(\varphi)$. Thus, f is isotopic to a self-orbit equivalence if and only if f_* preserves this set. Proposition 3.15 characterizes this set in terms of the vanishing of $d_\varphi(c, f)$. $\qquad\square$

Proof of Proposition 3.15   Fix $f = D_{\beta_1} \circ \cdots \circ D_{\beta_k}$, where D_{β_i} is a Dehn twist on T_i, and let $\tilde{f}$ be a lift of f to $\widetilde{M}$ chosen so that $\tilde{f}$ fixes some point $\tilde{x}$ on $\tilde{c}$. Recall that we assumed, without loss of generality, that all the T_i are quasitransverse.

Let $\gamma \in \pi_1(M)$ represent a periodic orbit c of φ. As a first trivial case, suppose c has no transverse intersections with any torus T_i, so it is either contained in a single torus or disjoint from all of them. Let $\tilde{c}$ be a lift of c. Note that c is isotopic to a curve c' disjoint from the support of f; lifting this isotopy means the lift $\tilde{c}$ is isotopic to a lift $\widetilde{c'}$ such that, after applying some deck transformation g of the cover, this is disjoint from the support of $\tilde{f}$. Thus, $\tilde{f}(g\tilde{c})$ is uniformly bounded distance away from $\tilde{f}(g\widetilde{c'}) = g\widetilde{c'}$, and this is uniformly bounded distance away from $g\tilde{c}$, showing that $g\tilde{c}$ (and hence also $\tilde{c}$) has bounded projection.

For the case where c intersects some T_i transversely, we may without loss of generality choose the product neighborhoods of the tori T_i in the definition of D_{β_i} small enough that, every time c crosses a torus T_i, it enters one side $T_i \times \{\pm 1\}$ and leaves the other, $T_i \times \{\mp 1\}$. Consider the positive-time ray $r = \{\varphi^t(\tilde{x}) : t \geq 0\} \subset \tilde{c}$, where $\tilde{x}$ is as before a point fixed by $\tilde{f}$. We will show that the projection of $\tilde{f}(r)$ to $\Lambda^s(\varphi)$ is bounded. Reversing the argument (using unstable leaves instead of stable) will show that the projection of $\tilde{f}(\tilde{c})$ to $\Lambda^s(\varphi)$ is also bounded, so we do only the forward case.

Between $\tilde{x}$ and $\gamma\tilde{x}$, the ray r intersects a finite number of lifts of the tori T_i on which the Dehn twists D_{β_i} are supported. Let $\mathcal{T}_1, \ldots, \mathcal{T}_n$ denote these lifts, indexed along the path of r so that r first intersects $\mathcal{T}_1$

after $\tilde{x}$. Note that two distinct lifts $\mathcal{T}_i$ and $\mathcal{T}_j$ may project to the same torus in M; this will happen whenever the orbit c crosses the same torus twice. Since the T_i are quasitransverse tori, each $\mathcal{T}_j$ projects to a string of lozenges $\mathscr{C}_j$. Let $L_0^{(1)}$ be the lozenge in $\mathscr{C}_1$ containing the projection of $\tilde{c}$. The main technical part of the proof is the following claim:

Claim 3.16 *Let r_0 denote the segment of r between $\tilde{x}$ and $\gamma\tilde{x}$.*

(1) *The stable leaf of $\tilde{f}(\gamma\tilde{x})$ intersects the lozenge $\eta^{d_\varphi(c,f)}(L_0^{(1)})$.*

(2) *There exists $N > 0$, depending only on f, such that the stable leaf saturation of $\tilde{f}(r_0)$ intersects the chain of lozenges $\mathscr{C}_1$ only between $\eta^{-N}(L_0^{(1)})$ and $\eta^N(L_0^{(1)})$.*

Given this claim, the proof of Proposition 3.15 can be finished quickly, by considering the positive iterates of r_0 under γ. Let us assume the claim for the moment and use it to derive the conclusion of Proposition 3.15.

We use the fact that $r = \bigcup_{i=1}^{\infty} \gamma^i(r_0)$. For the first direction, suppose that $d_\varphi(c, f) = 0$. Fix a segment $r_i = \gamma^i(r_0)$. Then $\gamma^{-i}(r_i) = r_0$, so, by Claim 3.16, the stable leaf of $\tilde{f}(r_i) = f_*(\gamma^i)\tilde{f}(r_0)$ intersects the lozenge $f_*(\gamma^i)(L_0^{(1)})$, and $\tilde{f}(r_i)$ intersects the chain of lozenges $\mathscr{C}_1$ only between $f_*(\gamma^i)\eta^{-N}(L_0^{(1)})$ and $f_*(\gamma^i)\eta^N(L_0^{(1)})$. But Observation 3.14 implies that $f_*(\gamma^i)(L_0^{(1)})$ is contained in the stable saturation of $L_0^{(1)}$; thus, $\tilde{f}(r_i)$ is contained in the union of the stable saturation of leaves between $\eta^{-N}(L_0^{(1)})$ and $\eta^N(L_0^{(1)})$ for all i; hence, r has bounded projection to $\Lambda^s(\varphi)$. As remarked above, applying the same argument to unstable leaves and using the negative time ray shows that c has bounded projection.

Conversely, if $\iota = i_\varphi(c, f) \neq 0$, then the argument above shows that $\widetilde{\mathscr{F}}^s(\tilde{f}(\gamma^n\tilde{x}))$ intersects $\eta^{n\iota}(L_1^{(1)})$; thus, the projection of $\tilde{f}(r)$ to $\Lambda^s(\varphi)$ is unbounded. $\qquad\square$

So, to finish the proof of the proposition, we only need to prove Claim 3.16.

Proof of Claim 3.16 Recall that f has support in disjoint small neighborhoods of the quasitransverse tori in M which lift to disjoint neighborhoods of $\mathcal{T}_j$ in $\widetilde{M}$, and we use $L_k^{(j)}$ to denote the string of lozenges associated to $\mathcal{T}_j$. Let $\tilde{N}_j$ denote the lifted neighborhood containing $\mathcal{T}_j$. We use these neighborhoods to split the ray r_0 into a union of intervals I_j and J_j, where $J_j := \tilde{N}_j \cap r_0$, and I_j denotes the connected component of r_0 between $\tilde{N}_{j-1}$ and $\tilde{N}_j$. The interval I_0 is the segment of r_0 from $\tilde{x}$ to $\tilde{N}_1$.

For each $\mathcal{T}_i$, let t_i denote the translation number of the Dehn twist on the corresponding torus. Notice that, since f is supported on the union of the projection of the neighborhoods $\tilde{N}_i$ to M, it follows that, for any j, there exists $h_j \in \pi_1(M)$ such that $\tilde{f}(I_j) = h_j(I_j) \subset h_j\tilde{c}$. Our choice of lift implies that h_0 is the identity.

To demonstrate the first statement of the claim, we will first establish the fact that, if $\tilde{f}(I_{i-1}) = h_{i-1}(I_{i-1})$ intersects a lozenge $h_{i-1}L_k^{(i)}$, then the stable leaf of $\tilde{f}(I_i)$ intersects the translate of $h_{i-1}L_k^{(i)}$ by $\eta^{\epsilon_i 2t_i}$.
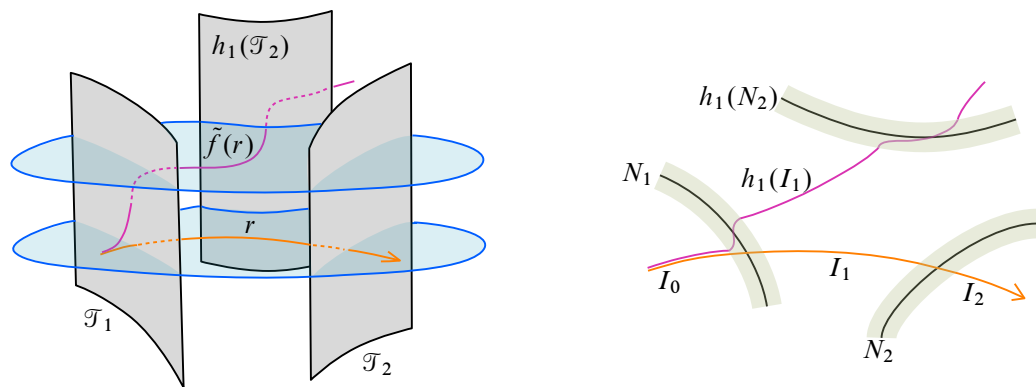
Figure 4: The action of the lift of a Dehn twist on a ray, shown in $\widetilde{M}$ on the left and schematically indicating intersections with lifted neighborhoods of tori on the right.

Here is the proof of the fact: Fix some i. The segment $\tilde{f}(J_i)$ is obtained from $h_{i-1}(J_i)$ by applying the (unique) lift of the Dehn twist supported on the projection of $\mathcal{T}_i$ that preserves $h_{i-1}\mathcal{T}_i$ and fixes $h_{i-1}(I_{i-1})$. The (projection to the orbit space of the) endpoint shared by $\tilde{f}(I_{i-1}) = h_{i-1}(I_{i-1})$ and $\tilde{f}(J_i)$ lies in some lozenge $L_k^{(i-1)}$ associated to $h_{i-2}(\mathcal{T}_{i-1})$. By Observation 3.14, the lozenge $h_{i-1}(L_{k'}^{(i)})$ for $h_{i-1}(\mathcal{T}_i)$ that also contains this point is such that $h_{i-1}(L_{k'}^{(i)})$ is contained in the saturation by stable leaves of $h_{i-2}(L_k^{(i-1)})$. Applying Lemma 3.12, we conclude that after applying the lifted Dehn twist to $h_{i-1}(J_i)$ (see Figure 4), the image of its other endpoint, which is also an endpoint of $\tilde{f}(I_i)$, lies in $\eta^{\epsilon_i 2t_i} h_{i-1}(L_{k'}^{(i)})$. This is contained in the saturation by stable leaves of $\eta^{\epsilon_i 2t_i} h_{i-2}(L_k^{(i-1)})$, which proves the fact.

To deduce (1) using the fact, we apply it iteratively, observing first that the stable leaf of $\tilde{f}(I_1)$ intersects $\eta^{\epsilon_1 2t_1}(L_0^{(1)})$ and some lozenge $h_1(L_k^{(2)})$, and the stable saturation of $L_k^{(2)}$ is contained in $\eta^{\epsilon_1 2t_1} L_0^{(1)}$. Thus, $\tilde{f}(I_2)$ has stable leaf intersecting $\eta^{\epsilon_2 2t_2} h_1(L_k^{(2)})$, and thus $\eta^{\epsilon_1 2t_1} \eta^{\epsilon_2 2t_2}(L_0^{(1)})$. Continuing iteratively, we conclude that $\tilde{f}(\gamma\tilde{x})$ has stable leaf intersecting the translate of $L_0^{(1)}$ by $\eta^{d_\varphi(c,f)}$.

Now the second part of the claim follows directly from the application of the second part of Lemma 3.12 in the proof above by choosing

$$N = \sum_{i=1}^{k} 2|t_i|. \qquad \qquad \square$$

3.2 Application: classification of self-orbit equivalences in special cases

We will now use the criterion given by Theorem 1.4 to describe the isotopy class of self-orbit equivalences in some special cases, starting with the proof of Corollaries 1.5 and 1.6.

Proof of Corollary 1.5 Suppose that D_α is a Dehn twist on a torus T in a direction α that is represented by a periodic orbit of φ. Then the translation number of α is zero, and thus, for any periodic orbit c of φ,

the displacement of c by D_α is zero. Therefore, the criterion of Theorem 1.4 implies that any such Dehn twist is isotopic to a self-orbit equivalence of φ.   □

Proof of Corollary 1.6 Let T be a quasitransverse torus in M and D_β a Dehn twist on T with nonzero translation number. Suppose first that T is separating. Consider a periodic orbit c of φ. Then either c does not intersect T or it intersects T an even number of times with alternating signs. In either case, $d_\varphi(c, D_\beta) = 0$. Thus, D_β is isotopic to a self-orbit equivalence by Theorem 1.4.

Now, if we assume instead that T is nonseparating, we claim that we can find a periodic orbit c of φ such that its intersections with T are always in the same transverse direction, and therefore are assigned the same sign. To do this, take a closed oriented loop based in T in M intersecting T exactly once, transversely, at the basepoint (and therefore inducing a transverse orientation of T). Lift this loop to a path in $\widetilde{M}$ with endpoints in lifts $\widetilde{T}_0$ and $\widetilde{T}_1$ of T, and iteratively choose successive lifts $\widetilde{T}_i$. Then, for each i, $\widetilde{T}_i$ separates $\widetilde{T}_{i-1}$ from $\widetilde{T}_{i+1}$, and no lift of T separates $\widetilde{T}_i$ from $\widetilde{T}_{i+1}$.

Fix n large, and let $\tilde{d}$ be a segment of an orbit in $\widetilde{M}$ with endpoints in $\widetilde{T}_0$ and $\widetilde{T}_n$. Such an orbit always exists because the flow is $\mathbb{R}$–covered and skew. If n is chosen large enough, $\tilde{d}$ projects down to an orbit segment in M that will contain points close enough to satisfy the conditions of the Anosov closing lemma. Hence, we obtain a periodic orbit c such that all its intersections with T have the same transverse orientation. We conclude that $d_\varphi(c, D_\beta)$ is nonzero, which, by the criterion, implies that D_β is not isotopic to a self-orbit equivalence.   □

The proof of this corollary gives an example of a general strategy to identify self-orbit equivalences by understanding the intersection of periodic orbits with quasitransverse tori. One could broaden this to give a characterization of self-orbit equivalences of a given flow in terms of the configuration of tori in M (the JSJ graph and geometry of the pieces) and translation numbers of periodic orbits in tori. Working through the details of this is beyond the scope of this article; instead we illustrate our criterion with some examples where we can virtually describe the group of self-orbit equivalences in the mapping class group.

For both of the following statements, we assume φ is a transversally orientable skew Anosov flow on M.

Theorem 3.17 *Suppose that the JSJ decomposition of M has a single piece which is atoroidal, glued to itself along n boundary tori $T_1, \ldots, T_n$. Let D_{per} be the group generated by Dehn twists on the T_i in the directions of the periodic orbits of φ.*

Let D be the finite-index subgroup of the mapping class group generated by Dehn twists on tori. Then D_{per} is the set of self-orbit equivalences in D.

Theorem 3.18 *Suppose that each torus of the JSJ decomposition of M is separating. Let D be the finite-index subgroup of the mapping class generated by Dehn twists on tori. Then the set of self-orbit equivalences in D is equal to the group generated by Dehn twists in the directions of periodic orbits inside Seifert pieces, together with any Dehn twists on the JSJ tori.*

Remark 3.19 Using Handel–Thurston surgery and Foulon–Hasselblatt Dehn surgeries on periodic orbits of geodesic flows, one can easily construct many examples of skew Anosov flows on manifolds as in the statements of Theorems 3.17 and 3.18.

The proofs of both statements use the fact that the group generated by Dehn twists on tori has finite index in the mapping class group of any orientable 3–manifold. (See [Johannson 1979] for an explanation and history of the proof.)

Proof of Theorem 3.17 By Corollary 1.5, any isotopy class $[f] \in D_{\text{per}}$ is represented by a self-orbit equivalence. For the other containment, suppose that f is a self-orbit equivalence contained in D. We assume for a contradiction that f cannot be written as a product of elements in D_{per}. Since elements of D_{per} are self-orbit equivalences, up to composition with such a map we can assume that $f = D_{\beta_1} \circ \cdots \circ D_{\beta_k}$, where $t_\varphi(\beta_i) \neq 0$ for all i. In fact, we will only use that one of these has nonzero translation number.

We now produce a periodic c such that $d(c, f) \neq 0$, leading to a contradiction by Theorem 1.4. Reindexing if needed, suppose T_1 is a torus in the support of f, with associated Dehn twist D_{β_1}. By hypothesis, T_1 is nonseparating in $M \smallsetminus (T_2 \cup \cdots \cup T_n)$. By considering lifts of curves as in the proof of Corollary 1.6, one can find a periodic orbit c of φ that intersects only T_1, and each intersection point has the same orientation. Thus, $d(c, f) = d(c, D_{\beta_1}) = \pm t_\varphi(\beta_1) \neq 0$, contradicting Theorem 1.4. □

Proof of Theorem 3.18 Given the assumption, all JSJ tori are separating, so Corollary 1.6 shows that any product of Dehn twists along them is isotopic to a self-orbit equivalence. This, together with Corollary 1.5, shows that any element of the group generated by Dehn twists in the direction of periodic orbits together with Dehn twists on the JSJ tori is isotopic to a self-orbit equivalence.

We are left to show that any element of the finite-index subgroup generated by Dehn twists that is a self-orbit equivalence is of this form. Let f be a self-orbit equivalence generated by Dehn twists; thus, f preserves each Seifert piece. Fix a Seifert piece S and let Σ denote the base orbifold of S, and g the restriction of f to Σ.

Up to composing f with some Dehn twists in the direction of periodic orbits on S and Dehn twists on the boundary of S, we can assume that g projects to the identity in the mapping class group of the base orbifold Σ (see [Johannson 1979] or [Bonatti et al. 2020, Section 4.7]). Thus, for any $\gamma \in \pi_1(S)$ that is obtained as a lift of an element of $\pi_1(\Sigma)$ that is not homotopic to a boundary component of Σ, there exists $k \in \mathbb{Z}$ such that $g_*(\gamma) = \gamma s^k$, where s is the element of $\pi_1(S)$ representing a regular fiber of the Seifert fibration of S.

For any such γ, the group $H = \langle \gamma, s \rangle$ is a subgroup of $\pi_1(M)$ homeomorphic to $\mathbb{Z}^2$, and hence there exists a (unique) string of lozenges $\mathscr{C}$ in $\mathbb{O}$ that is preserved by H. The permutation action of H on this string gives a homomorphism $H \to \mathbb{Z}$; thus, there exists some nontrivial $\gamma' \in H$ whose action on $\mathbb{O}$ fixes

each corner of $\mathscr{C}$; equivalently, the conjugacy class of γ' is represented by the periodic orbits of φ that project to the corners of $\mathscr{C}$.

Since the weak foliations are transverse to the Seifert fibration, s acts as a translation on the corners of $\mathscr{C}$ (see [Barbot and Fenley 2013]). In particular, if $k \neq 0$, then $g_*(\gamma') = \gamma' s^k$ does not fix a point of $\mathbb{O}$, and hence is not represented by a periodic orbit of φ. This contradicts the fact that g was the restriction of a self-orbit equivalence. Thus, $k = 0$, and, since the choice of γ was arbitrary, we conclude that g is isotopic to the identity.

Applying this argument to the restriction of f to each Seifert piece, we conclude that f is a product of Dehn twists in the direction of periodic orbits and Dehn twists on the JSJ tori, as claimed. $\qquad\square$

Remark 3.20 The proof of Theorem 3.18 gives a virtual characterization of the self-orbit equivalences of a given flow that fix a Seifert piece and are isotopic to identity on its boundary: they are exactly the isotopy classes of the group generated by Dehn twists in the direction of periodic orbits in the Seifert piece.

4 Application to contact flows

We will now apply some results of contact geometry to obtain the reverse direction of Theorem 1.12, as well as its corollaries. Recall that a (coorientable) contact structure ξ is *Anosov* if ξ admits an Anosov Reeb flow, ie there exists a 1–form α such that $\xi = \ker \alpha$ and the Reeb flow of α is Anosov.

We will use as a black box the theory of cylindrical contact homology. Introduced in a more general context by Eliashberg, Givental and Hofer [Eliashberg et al. 2000], it has been proven to be well defined for dynamically convex contact structures by Hutchings and Nelson [2016]. This context includes the case of Anosov contact forms: if the Reeb flow of a contact 1–form α is Anosov, then it is automatically nondegenerate and, since Anosov flows have no contractible orbits, α is dynamically convex (see also eg [Macarini and Paternain 2012]). Some of the fundamental results in this theory can be summarized as follows:

Theorem 4.1 *If α_1 and α_2 are nondegenerate dynamically convex contact forms on a closed, hypertight, contact 3–manifold (M, ξ), then $C\mathbb{H}_{\mathrm{cyl}}^{\Lambda}(\alpha_1) \cong C\mathbb{H}_{\mathrm{cyl}}^{\Lambda}(\alpha_2)$ for any set Λ of free homotopy classes in M. The space $C\mathbb{H}_{\mathrm{cyl}}^{\Lambda}(\alpha_i)$ is a $\mathbb{Q}$–vector space, it is the homology of a complex generated by the periodic orbits of the Reeb flow of α_i in the free homotopy classes belonging to Λ.*

Having the cylindrical contact homology groups associated to α_1 and α_2 being isomorphic is not quite enough for our purpose: what we will need in order to apply Theorem 1.2 is to obtain that the chain complexes themselves are equal, when α_1 and α_2 are Anosov contact forms. This follows from the fact that the differential in the chain complex is trivial. That result was proven by Macarini and Paternain [2012, Theorem 2.1] for any Anosov contact structure (and independently by Vaugon — see [Foulon et al. 2021] — with the additional assumption that the foliations are orientable).

Proof of Theorem 1.12, reverse implication   Let φ_1 and φ_2 be two contact Anosov flows with respective contact forms α_1 and α_2 and contact structures $\xi_1 = \ker \alpha_1$ and $\xi_2 = \ker \alpha_2$. Suppose that ξ_1 and ξ_2 are contactomorphic. Then there exists a diffeomorphism $g \colon M \to M$ such that $g_* \xi_1 = \xi_2$.

In our situation, the 1–forms $g^* \alpha_1$ and α_2 are two contact forms of (M, ξ_2) with Anosov Reeb flows. Applying the theorem on cylindrical contact homology above, where Λ runs over all possible free homotopy classes in $\pi_1(M)$, we deduce that each homotopy class represented by a periodic orbit of the Reeb flow of $g^* \alpha_1$ is also represented by a periodic orbit of the Reeb flow φ_2.

The Reeb flow of $g^* \alpha_1$ is $g^{-1} \circ \varphi_1^t \circ g$. Thus, by Theorem 1.2, $g^{-1} \circ \varphi_1 \circ g$ and φ_2 are isotopically equivalent. Equivalently, φ_1 and φ_2 are orbit equivalent. If additionally g is isotopic to the identity, we conclude that φ_1 and φ_2 are isotopically equivalent. $\square$

Using this direction of Theorem 1.12, together with the coarse classification of contact structures in dimension 3 of Colin–Giroux–Honda, we can quickly deduce Theorem 1.13:

Proof of Theorem 1.13   By [Colin et al. 2009, théorème 6], on any irreducible 3–manifold, there exist at most finitely many noncontactomorphic contact structures with bounded Giroux torsion. Now Proposition A.1 shows that the contact structure of any contact Anosov flow has zero Giroux torsion. Thus, Theorem 1.12 implies the result.[4] $\square$

Appendix   Further applications to contact topology
joint with Jonathan Bowden

In this appendix, we show that any Anosov contact structure has zero Giroux torsion, prove the converse to Theorem 1.12, and then use this to obtain new results about contact structures.

Proposition A.1   *Let ξ be an Anosov contact structure on a 3–manifold M. Then ξ has zero Giroux torsion. In fact, a double cover of M will be strongly symplectically fillable.*

Proof   Let α be a contact 1–form such that $\ker \alpha = \xi$ and such that the Reeb flow φ of α is Anosov. Let X be the Reeb vector field. Consider the 4–manifold $M \times [-1, 1]$ with symplectic form $\omega = d(e^t \alpha)$, where t is the coordinate on $[-1, 1]$. Then $\xi = \ker \alpha$ is a contact structure that we can assume without loss of generality to be positive and, since $\omega|_{M \times \{1\}} = e^1 \, d\alpha$, the form ω is nonzero on ξ.

Up to taking a double cover, we can assume that φ is transversally orientable, ie its Anosov splitting is orientable.

Let X^{ss} and X^{uu} be vector fields in, respectively, the stable and unstable direction of the Anosov splitting of X, with orientations chosen so that the plane spanned by $X^{\mathrm{ss}} + X^{\mathrm{uu}}$ and X defines a (coorientable) contact structure ξ_- with *negative* orientation (see [Mitsumatsu 1995]). Put this contact structure on $M \times \{-1\}$.

---

[4]When M is hyperbolic, one does not need Bowden's result, thanks to [Colin et al. 2009, théorème 2].

Since the defining property of a contact form is open in the C^1 topology, we can take a C^1–small approximation $\widetilde{\xi}_-$ of ξ_- such that X is transverse to $\widetilde{\xi}_-$. Then ω is nonzero on $\widetilde{\xi}_-$ (since $\omega|_{M\times\{-1\}} = e^{-1}\,d\alpha$, so its kernel is spanned by X).

This gives us a weak semifilling of $(M\times\{1\},\xi)$. By [Eliashberg 2004, Corollary 1.4], a weak semifilling can be capped off to give a weak filling. Since ω is exact, by [Eliashberg 2004, Proposition 4.1] this weak filling can be modified to give a strong filling (this result was independently obtained by Etnyre [2004]). Now, Gay [2006, Corollary 3] proved that any contact plane field that is strongly fillable has zero Giroux torsion. This proves the proposition for possibly a double cover of M.

Now the Giroux torsion of a contact structure cannot decrease under finite covers, since any component of a finite cover of a Giroux torsion domain is again a Giroux torsion domain. So, in any case, a contact structure ξ has zero Giroux torsion if and only if any finite lift of it has zero Giroux torsion. This proves the proposition for the original ξ. $\qquad\square$

Thanks to Giroux's correspondence [2002] between open books and contact structures, we can prove the following, giving the second implication needed for the statement of Theorem 1.12:

Theorem A.2 *Let φ_1 and φ_2 be two contact Anosov flows with respective contact structures ξ_1 and ξ_2. If φ_1 and φ_2 are orbit equivalent (resp. isotopically equivalent) then ξ_1 and ξ_2 are contactomorphic (resp. isotopic).*

To apply Giroux's result in our proof, we first need to recall a result about Birkhoff sections, starting with their definition:

Definition A.3 Let φ be a flow on a 3–manifold M. A surface S is called a *topological Birkhoff section* of φ if

   (i) S is topologically immersed in M, and its interior is topologically embedded;

  (ii) the flow φ is topologically transverse to the interior of S;

 (iii) each connected component of the boundary of S consists of a periodic orbit of φ;

  (iv) every orbit of φ intersects S; and

   (v) the return time of φ to the interior of S is uniformly bounded above and away from zero.

A topological Birkhoff section that is smoothly immersed is called a (smooth) Birkhoff section.

A Birkhoff section S in an orientable manifold is called *positive* if the orientations of all the boundary orbits correspond to the orientation induced by the flow on the interior of *S*.

Fried showed that any transitive Anosov flow has a Birkhoff section, which can in fact be taken to be embedded on the boundary as well. Moreover, Bonatti and Guelman [2010] show that this section can be assumed to be *tame*, meaning that, after an isotopy along flow lines, one can assume that the restriction of the Birkhoff section to a small tubular neighborhood of any component of its boundary is a smooth

helicoid. Finally, Marty [2021] showed that any $\mathbb{R}$–covered Anosov flow admits a *positive* tame Birkhoff section. Moreover, an adaptation of Marty's proof can be seen to yield a section that is also embedded on the boundary (T Marty, personal communication, 2021).[5] Thus, one obtains an open book supporting the contact structure. We recall:

Definition A.4 (open book) Let $B \subset M$ be an oriented link in a connected oriented manifold. Then an open book (B, θ) with binding B is a fibration with connected oriented (noncompact) fibers $\theta : M \smallsetminus B \to S^1$ such that, in a neighborhood of each binding component, the map is equivalent to the map given by projecting to the angular polar coordinate on $(D^2 \smallsetminus \{0\}) \times S^1$, and the boundary of any fiber agrees with B as an oriented link.

The fibers of an open book are called *pages*. Note that the tameness condition in [Bonatti and Guelman 2010] corresponds precisely to a Birkhoff section S inducing an open book with binding the (oriented) periodic orbits ∂S.

Definition A.5 (supporting open book) An open book (B, θ) supports a (cooriented) contact structure (M, ξ) if there is a contact form α for ξ such that:

- The form $d\alpha$ is positive on the pages of θ, which are oriented to be compatible with the binding.
- The form α is positive on (each component of) B.

The fundamental fact due to Giroux [2002] is that any two contact structures supported by a *fixed* open book are isotopic through contact structures supported by the open book. This is essentially due to the fact that the above condition is convex, although some care is needed near the binding.

In order to apply the above, we will use the following result of [Bonatti and Guelman 2010]:

Lemma A.6 [Bonatti and Guelman 2010, Lemma 4.16] *Let S be a topological Birkhoff section of a flow φ; then S can be isotoped along the orbits of φ to a smooth tame Birkhoff section S'.*

Proof of Theorem A.2 Let h be an orbit equivalence between φ_1 and φ_2. Up to conjugating φ_1 by a diffeomorphism in the isotopy class of h, we can assume that φ_1 and φ_2 are isotopically equivalent and h is isotopic to the identity. Showing that ξ_1 and ξ_2 are isotopic for this new flow will imply that ξ_1 and ξ_2 are contactomorphic for the original one.

Let S be a smooth Birkhoff section for φ_1. Then $h(S)$ is a topological Birkhoff section for φ_2. By the lemma above, we can isotope $h(S)$ to a smooth tame Birkhoff section S_2 of φ_2. Thus, the open book (B_2, θ_2) induced by S_2 supports ξ_2. Now we isotope h to a smooth map g relative to the boundary of the

---

[5]One could also run the proof below using *rational* open books, which correspond to positive Birkhoff sections with immersed boundaries; see [Baker et al. 2012].

original Birkhoff section S. Then the open book (B_2, θ_2) pulls back to an open book (B', θ') with page $S' = g^{-1} S_2$ that supports $\xi'_1 = g^* \xi_2$. This smoothing can be arranged so that preimages of pages agree with those of an open book (B, θ) coming from the original Birkhoff section S near the binding. Then one notes that the open book (B', θ') is isotopic to the original one and we deduce that the corresponding contact structures are contactomorphic. Since h (and hence g) is isotopic to the identity, they are in fact isotopic. $\square$

A.1 Applications to contact topology

Now we can use Theorem 1.12 to translate to the language of contact topology some known results about Anosov flows. As a first example, Theorem 1.12 implies that the examples of skew Anosov flows on hyperbolic 3–manifolds built in [Bowden and Mann 2022] (which are all contact flows when done using Foulon–Hasselblatt contact surgery) have noncontactomorphic contact structures. Thus, we immediately obtain:

Theorem A.7 *For any $N \in \mathbb{N}$ there exists an hyperbolic 3–manifold with at least N noncontactomorphic Anosov contact structures.*

This result answers affirmatively a question raised in [Foulon et al. 2021]; see also [Bowden and Mann 2022, Question 7.4].

We can also translate Theorems 3.17 and 3.18 to a description of contact transformation groups, as follows. For a 3–manifold M with Anosov contact structure ξ, we follow [Giroux and Massot 2017] and denote by $\mathscr{D}(M, \xi)$ the group of contact transformations of (M, ξ). Thus, there is a natural inclusion of $\pi_0 \mathscr{D}(M, \xi)$ in $\mathrm{MCG}(M)$.

Theorem A.8 *Let ξ be a Anosov contact structure on M. Suppose that either*

(1) *M has a unique JSJ piece which is atoroidal, or*

(2) *each torus of the JSJ decomposition of the manifold M is separating.*

In the first case, let D_ξ denote the subgroup of $\mathrm{MCG}(M)$ generated by Dehn twists on the JSJ tori. In the second case, let D_ξ be the subgroup of $\mathrm{MCG}(M)$ generated by all Dehn twists in the directions of periodic orbits together with any Dehn twists on the JSJ tori.

Then any class $[f] \in D_\xi$ admits a representative $f \in \mathscr{D}(M, \xi)$, and, conversely, there exists $n \in \mathbb{N}$ such that, for any class $[f] \in \pi_0 \mathscr{D}(M, \xi)$, we have $[f]^n \in D_\xi$.

Remark A.9 This result partially extends a theorem of Giroux and Massot [2017], who obtained this for the case of Seifert fibered manifolds. Note that the result of Giroux and Massot is more precise, as ours only gives a description of $\pi_0 \mathscr{D}(M, \xi)$ up to finite powers.

Remark A.10 It is not necessary to know the Anosov Reeb flow in order to detect which Dehn twist is in a direction of a periodic orbit, by the following observation: Let T be an embedded torus that is quasitransverse to the Anosov flow φ. Up to an arbitrarily small perturbation, one can put T in a convex position with respect to the contact structure ξ. Then an element $\alpha \in \pi_1(T)$ corresponding to a periodic orbit of the flow φ also corresponds to the free homotopy class of a connected component of the dividing set of the characteristic foliation of ξ on T. Therefore, one can use Theorem A.8 (or the translation of Corollary 1.5, which can be obtained in the same way) directly in contact geometry without having to go through the Anosov side.

Proof of Theorem A.8 We start by proving the converse implication. Let $f \in \mathscr{D}(M, \xi)$ and let φ be the (Anosov) Reeb flow. Then $f^{-1} \circ \varphi^t \circ f$ is a contact Anosov flow with contact structure $f_* \xi = \xi$. Thus, by Theorem 1.12, $f^{-1} \circ \varphi^t \circ f$ and φ^t are isotopically equivalent. Let $h \colon M \to M$ be an orbit equivalence isotopic to the identity between $f^{-1} \circ \varphi^t \circ f$ and φ^t; then $h \circ f$ is a self-orbit equivalence of φ^t. Hence, by Theorems 3.17 or 3.18 depending on the case, there exists n such that $[f^n] = [(h \circ f)^n] \in D_\xi$.

Now, for the second part, let $[f]$ be a class in D_ξ. Then, by Theorems 3.17 or 3.18, $\psi = f^{-1} \circ \varphi \circ f$ is isotopically equivalent to φ. Moreover, the contact structure of ψ is $f_* \xi$, where ξ is the contact structure of φ. By Theorem 1.12, $f_* \xi$ and ξ are isotopic, so there exists g in the same isotopy class as f such that $g_* \xi = \xi$. That is, $g \in \mathscr{D}(M, \xi)$. $\qquad \square$

References

[Baker et al. 2012] **K L Baker**, **J B Etnyre**, **J Van Horn-Morris**, *Cabling, contact structures and mapping class monoids*, J. Differential Geom. 90 (2012) 1–80 MR Zbl

[Barbot 1995a] **T Barbot**, *Caractérisation des flots d'Anosov en dimension 3 par leurs feuilletages faibles*, Ergodic Theory Dynam. Systems 15 (1995) 247–270 MR Zbl

[Barbot 1995b] **T Barbot**, *Mise en position optimale de tores par rapport à un flot d'Anosov*, Comment. Math. Helv. 70 (1995) 113–160 MR Zbl

[Barbot 2001] **T Barbot**, *Plane affine geometry and Anosov flows*, Ann. Sci. École Norm. Sup. 34 (2001) 871–889 MR Zbl

[Barbot 2005] **T Barbot**, *De l'hyperbolique au globalement hyperbolique*, Habilitation à diriger des recherches, Université Claude Bernard Lyon 1 (2005) Available at https://theses.hal.science/tel-00011278

[Barbot and Fenley 2013] **T Barbot**, **S R Fenley**, *Pseudo-Anosov flows in toroidal manifolds*, Geom. Topol. 17 (2013) 1877–1954 MR Zbl

[Barthelmé and Fenley 2017] **T Barthelmé**, **S R Fenley**, *Counting periodic orbits of Anosov flows in free homotopy classes*, Comment. Math. Helv. 92 (2017) 641–714 MR Zbl

[Barthelmé and Gogolev 2019] **T Barthelmé**, **A Gogolev**, *A note on self orbit equivalences of Anosov flows and bundles with fiberwise Anosov flows*, Math. Res. Lett. 26 (2019) 711–728 MR Zbl

[Barthelmé et al. 2023] **T Barthelmé**, **S R Fenley**, **R Potrie**, *Collapsed Anosov flows and self orbit equivalences*, Comment. Math. Helv. 98 (2023) 771–875 MR Zbl

[Bonatti and Guelman 2010]   **C Bonatti**, **N Guelman**, *Axiom A diffeomorphisms derived from Anosov flows*, J. Mod. Dyn. 4 (2010) 1–63   MR   Zbl

[Bonatti and Iakovoglou 2023]   **C Bonatti**, **I Iakovoglou**, *Anosov flows on* 3*–manifolds: the surgeries and the foliations*, Ergodic Theory Dynam. Systems 43 (2023) 1129–1188   MR   Zbl

[Bonatti et al. 2020]   **C Bonatti**, **A Gogolev**, **A Hammerlindl**, **R Potrie**, *Anomalous partially hyperbolic diffeomorphisms, III: Abundance and incoherence*, Geom. Topol. 24 (2020) 1751–1790   MR   Zbl

[Bowden and Mann 2022]   **J Bowden**, **K Mann**, *C^0 stability of boundary actions and inequivalent Anosov flows*, Ann. Sci. École Norm. Sup. 55 (2022) 1003–1046   MR   Zbl

[Casson and Jungreis 1994]   **A Casson**, **D Jungreis**, *Convergence groups and Seifert fibered* 3*–manifolds*, Invent. Math. 118 (1994) 441–456   MR   Zbl

[Colin et al. 2009]   **V Colin**, **E Giroux**, **K Honda**, *Finitude homotopique et isotopique des structures de contact tendues*, Publ. Math. Inst. Hautes Études Sci. 109 (2009) 245–293   MR   Zbl

[Eliashberg 2004]   **Y Eliashberg**, *A few remarks about symplectic filling*, Geom. Topol. 8 (2004) 277–293   MR   Zbl

[Eliashberg et al. 2000]   **Y Eliashberg**, **A Givental**, **H Hofer**, *Introduction to symplectic field theory*, Geom. Funct. Anal. special volume (2000) 560–673   MR   Zbl

[Etnyre 2004]   **J B Etnyre**, *On symplectic fillings*, Algebr. Geom. Topol. 4 (2004) 73–80   MR   Zbl

[Farb and Franks 2003]   **B Farb**, **J Franks**, *Group actions on one-manifolds, II: Extensions of Hölder's theorem*, Trans. Amer. Math. Soc. 355 (2003) 4385–4396   MR   Zbl

[Fenley 1994]   **S R Fenley**, *Anosov flows in* 3*–manifolds*, Ann. of Math. 139 (1994) 79–115   MR   Zbl

[Fisher and Hasselblatt 2019]   **T Fisher**, **B Hasselblatt**, *Hyperbolic flows*, Eur. Math. Soc., Berlin (2019)   MR   Zbl

[Foulon and Hasselblatt 2013]   **P Foulon**, **B Hasselblatt**, *Contact Anosov flows on hyperbolic* 3*–manifolds*, Geom. Topol. 17 (2013) 1225–1252   MR   Zbl

[Foulon et al. 2021]   **P Foulon**, **B Hasselblatt**, **A Vaugon**, *Orbit growth of contact structures after surgery*, Ann. H. Lebesgue 4 (2021) 1103–1141   MR   Zbl

[Funar 2013]   **L Funar**, *Torus bundles not distinguished by TQFT invariants*, Geom. Topol. 17 (2013) 2289–2344   MR   Zbl

[Gabai 1992]   **D Gabai**, *Convergence groups are Fuchsian groups*, Ann. of Math. 136 (1992) 447–510   MR   Zbl

[Gay 2006]   **D T Gay**, *Four-dimensional symplectic cobordisms containing three-handles*, Geom. Topol. 10 (2006) 1749–1759   MR   Zbl

[Giroux 2002]   **E Giroux**, *Géométrie de contact: de la dimension trois vers les dimensions supérieures*, from "Proceedings of the International Congress of Mathematicians, II" (T Li, editor), Higher Ed. Press, Beijing (2002) 405–414   MR   Zbl

[Giroux and Massot 2017]   **E Giroux**, **P Massot**, *On the contact mapping class group of Legendrian circle bundles*, Compos. Math. 153 (2017) 294–312   MR   Zbl

[Hutchings and Nelson 2016]   **M Hutchings**, **J Nelson**, *Cylindrical contact homology for dynamically convex contact forms in three dimensions*, J. Symplectic Geom. 14 (2016) 983–1012   MR   Zbl

[Johannson 1979]   **K Johannson**, *Homotopy equivalences of* 3*–manifolds with boundaries*, Lecture Notes in Math. 761, Springer (1979)   MR   Zbl

[Kovačević 1999]   **N Kovačević**, *Möbius-like groups of homeomorphisms of the circle*, Trans. Amer. Math. Soc. 351 (1999) 4791–4822   MR   Zbl

[Liverani 2004]   **C Liverani**, *On contact Anosov flows*, Ann. of Math. 159 (2004) 1275–1312   MR   Zbl

[Macarini and Paternain 2012]   **L Macarini**, **G P Paternain**, *Equivariant symplectic homology of Anosov contact structures*, Bull. Braz. Math. Soc. 43 (2012) 513–527   MR   Zbl

[Marty 2021]   **T Marty**, *Flots d'Anosov et sections de Birkhoff*, PhD thesis, Université Grenoble Alpes (2021) Available at https://theses.hal.science/tel-03510071

[Mitsumatsu 1995]   **Y Mitsumatsu**, *Anosov flows and non-Stein symplectic manifolds*, Ann. Inst. Fourier (Grenoble) 45 (1995) 1407–1421   MR   Zbl

[Plante 1981]   **J F Plante**, *Anosov flows, transversely affine foliations, and a conjecture of Verjovsky*, J. Lond. Math. Soc. 23 (1981) 359–362   MR   Zbl

[Smale 1998]   **S Smale**, *Mathematical problems for the next century*, Math. Intelligencer 20 (1998) 7–15   MR   Zbl

[Thurston 1997]   **W P Thurston**, *Three-manifolds*, *foliations and circles*, *I*, preprint (1997)   arXiv math/9712268 Reprinted in "Collected works with commentary, I: Foliations, surfaces and differential geometry" (B Farb, D Gabai, S P Kerckhoff, editors), Amer. Math. Soc., Providence, RI (2022) 353–412

Department of Mathematics and Statistics, Queen's University
Kingston ON, Canada

Department of Mathematics, Cornell University
Ithaca, NY, United States

Fakultät für Mathematik, Universität Regensburg
Regensburg, Germany

thomas.barthelme@queensu.ca,   k.mann@cornell.edu,
jonathan.bowden@mathematik.uni-regensburg.de

sites.google.com/site/thomasbarthelme,   https://e.math.cornell.edu/people/mann

Microlocal theory of Legendrian links and cluster algebras

ROGER CASALS

DAPING WENG

We show the existence of quasicluster $\mathscr{A}$–structures and cluster Poisson structures on moduli stacks of sheaves with singular support in the alternating strand diagram of grid plabic graphs by studying the microlocal parallel transport of sheaf quantizations of Lagrangian fillings of Legendrian links. The construction is in terms of contact and symplectic topology, showing that there exists an initial seed associated to a canonical relative Lagrangian skeleton. In particular, mutable cluster $\mathscr{A}$–variables are intrinsically characterized via the symplectic topology of Lagrangian fillings in terms of dually $\mathbb{L}$–compressible cycles. New ingredients are introduced throughout, including the initial weave associated to a grid plabic graph, cluster mutation along nonsquare faces of a plabic graph, possibly including lollipops, the concept of sugar-free hull, and the notion of microlocal merodromy. Finally, we prove the existence of the cluster DT transformation for shuffle graphs, constructing a contact-geometric realization and an explicit reddening sequence, and establish cluster duality for the cluster ensembles.

13F60, 53D12

1 Introduction

The object of this article will be to show the existence of intrinsically symplectic quasicluster K_2–structures and quasicluster Poisson structures on moduli stacks of sheaves with singular support in the alternating strand diagram of a complete grid plabic graph. The construction of such quasicluster structures is achieved via contact and symplectic topology, based on the recently developed machinery of Legendrian weaves, and we show that there exists a canonical initial quasicluster seed associated to a relative Lagrangian skeleton. This is the first manuscript proving the existence of such cluster structures for these general moduli stacks, and entirely in symplectic geometric terms, as well as introducing the first symplectic topological definition of cluster $\mathscr{A}$–variables associated to Lagrangian fillings of Legendrian links. In
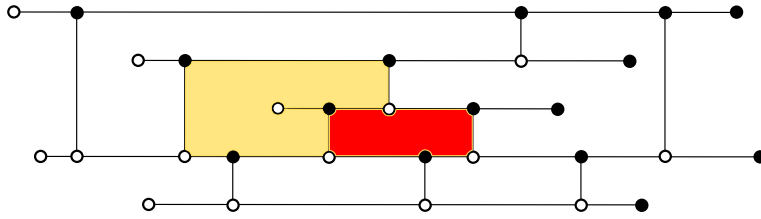
Figure 1: The quasicluster K_2–structure we construct for this grid plabic graph is on the coordinate ring of the moduli of decorated sheaves on $\mathbb{R}^2$ with singular support in a max-tb Legendrian representative of the $m(9_6)$ knot.

particular, our constructions admit natural contact and symplectic invariance and functoriality properties, and the cluster variables can be named and computed after performing Hamiltonian isotopies.

Several new ingredients are introduced for this purpose, among them the initial weave of a grid plabic graph, cluster mutations along nonsquare faces, possibly with lollipops, the concept of sugar-free hulls, and the notion of microlocal merodromy. Microlocal merodromies capture microlocal parallel transport along a relative cycle and they are crucial in defining a set of initial cluster $\mathscr{A}$–variables. From a contact geometry viewpoint, embedded Lagrangian disks whose boundaries lie on embedded exact Lagrangian fillings have a central role. This allows for geometric characterizations of mutable and frozen vertices, which arise from relative homology groups of triples, and naturally explains the appearance of quasicluster structures.

1.1 Scientific context

Cluster algebras, first introduced by S Fomin and A Zelevinsky [2002; 2003; Berenstein et al. 2005] in the context of Lie theory, are commutative rings endowed with a set of distinguished generators that have remarkable combinatorial structures. Cluster varieties, a geometric enrichment of cluster algebras introduced by V Fock and A Goncharov [2006b; 2006a], are affine schemes equipped with an atlas of torus charts whose transition maps obey certain combinatorial rules. Cluster varieties come in dual pairs consisting of a cluster K_2–variety, also known as a cluster $\mathscr{A}$–variety, and a cluster Poisson variety, also known as a cluster $\mathscr{X}$–variety. In particular, the coordinate ring of a cluster $\mathscr{A}$–variety coincides with an upper cluster algebra; see Berenstein, Fomin and Zelevinsky [Berenstein et al. 2005].

Since their introduction, cluster algebras and cluster varieties have appeared in many contexts, such as Teichmüller theory [Fock and Goncharov 2006b; Fomin et al. 2008; Gekhtman et al. 2005], birational geometry [Gross et al. 2015; 2018; Hacking and Keel 2018], the Riemann–Hilbert correspondence [Allegretti 2021; Neitzke 2014; Gaiotto et al. 2010], exact WKB analysis [Iwaki and Nakanishi 2014; 2016], and the study of positroid and Richardson varieties [Galashin and Lam 2023; Serhiyenko et al. 2019]. The first appearance of cluster mutations in symplectic geometry occurred in the study of wall-crossing formulas, following the work of D Auroux, K Fukaya, M Kontsevich, P Seidel, Y Soibelman and others; see eg [Auroux 2007; 2009; Pascaleff and Tonkonog 2020]. We also thank Goncharov for pointing out to us his recent work with Kontsevich [2021] focusing on noncommutative clusters, which also aligns well with the

developments we present here. The first hint that cluster $\mathscr{X}$–structures might naturally exist in the symplectic study of Legendrian knots was provided in [Shende et al. 2019], where it was computed how certain absolute monodromies around a square plabic face change under a square move in a plabic fence. See also the generalization presented in [Shende et al. 2016]. In conjunction, [Shende et al. 2019; 2016] should imply the existence of partial $\mathscr{X}$–structures for certain moduli stacks of sheaves singularly supported in the Legendrian lifts of the alternating strand diagrams of plabic fences. Nevertheless, they do not imply the existence of the full cluster $\mathscr{X}$–structures, nor the full cluster $\mathscr{A}$–structures and certainly not the fact that the rings of regular functions are cluster algebras. (See Section 2.8.) These stronger statements are proven here.

There are two obstacles to proving the existence of a cluster $\mathscr{A}$–structure. First, many plabic faces are typically not square and may contain lollipops; thus, one needs a new construction that both associates a cluster variable to them and allows for a geometric mutation to be performed. Second, more fundamental, is the regularity problem: even if all faces are square, the absolute monodromies are *not* global regular functions, and it is not possible to deduce the existence of a cluster structure purely from these microlocal monodromies. These obstacles are unavoidable if one is restricted to either plabic graphs or absolute cycles, both of which are limiting constraints in that approach.

Our new approach uses Legendrian weaves, which are more versatile than plabic graphs, and actually builds cluster $\mathscr{A}$–variables from relative cycles, which is stronger than the absolute analogue; see Section 2.8. In particular, we overcome both obstacles above, resolving the regularity problem, and finally prove the existence of cluster $\mathscr{A}$–structures and, consequently, cluster $\mathscr{X}$–structures in entirely symplectic topological terms. Some of our previous work used ideas from the theory of cluster algebras for new applications to contact and symplectic geometry — see eg [Casals 2022; 2020; 2021; Casals and Zaslow 2022; Gao et al. 2020a] — including the discovery of infinitely many Lagrangian fillings for many Legendrian links [Casals and Gao 2022]. This article builds in the converse direction, using contact and symplectic topology to construct (upper) cluster algebras, and symplectic topological results to deduce algebraic properties. In fact, we also know $\mathscr{A} = \mathscr{U}$ by [Casals et al. 2022], which builds on the present manuscript.

Note that what can be deduced from [Casals et al. 2020; Casals and Gao 2022; Casals and Zaslow 2022; Gao et al. 2020a; 2020b] is that certain moduli spaces that appear in contact topology are sometimes abstractly isomorphic to certain affine varieties, which themselves can independently be endowed[1] with cluster structures, but currently there does not exist any symplectic construction or characterization of cluster $\mathscr{A}$–variables or general cluster $\mathscr{X}$–variables, nor a symplectic geometric proof of the existence of cluster structures on these moduli spaces, nor even a geometric understanding of frozen variables. In particular, none of these previous constructions is known to have any Hamiltonian or Legendrian invariance properties, which are crucial in contact and symplectic topology. In fact, in all previous constructions even the initial seeds cannot be named after a Hamiltonian isotopy (eg even after a Reidemeister I or II move) and no symplectic computation or interpretation of cluster $\mathscr{A}$–variables existed. This work finally

---

[1]Explicitly, double Bott–Samelson cells for [Gao et al. 2020a], and positroids for [Casals and Gao 2022; Shende et al. 2019]. These instances are, in any case, particular cases of the moduli stacks that we associate to grid plabic graphs.

resolves this matter and, as we shall see, interesting symplectic features appear with regards to both mutable and frozen variables.

1.2 Main results

Let $\Lambda \subset (T^*_\infty \mathbb{R}^2, \xi_{st})$ be a Legendrian link in the ideal contact boundary of the cotangent bundle of the plane $\mathbb{R}^2$, and $T \subset \Lambda$ a set of marked points. The precise details and definitions for these contact-geometric objects are provided in Section 2. Let $L \subset (T^*\mathbb{R}^2, \lambda_{st})$ be an embedded exact Lagrangian filling of Λ. By definition, an embedded closed curve $\gamma \subset L$ is said to be $\mathbb{L}$–compressible if there exists a properly embedded Lagrangian 2–disk $D \subset (T^*\mathbb{R}^2 \setminus L)$ such that $\partial \overline{D} \cap L = \gamma \subset \mathbb{R}^4$. A collection $\{\gamma_1, \ldots, \gamma_\ell\}$ of such curves, with a choice of $\mathbb{L}$–compressing disk for each curve, is said to be an $\mathbb{L}$–compressing system for L if the curves form a maximal linearly independent subset in $H_1(L)$. In line with this, we will use Lagrangian disk surgeries, as defined in [Polterovich 1991; Yau 2017].

Consider also the moduli stack $\mathfrak{M}(\Lambda, T)$ of decorated microlocal rank-one constructible sheaves on $\mathbb{R}^2$ with singular support contained in Λ, as defined in Section 2.7.3, following [Kashiwara and Schapira 1990; Guillermou et al. 2012], which is invariant under contact isotopies. Let $\mathbb{G} \subset \mathbb{R}^2$ be a complete grid plabic graph and $\Lambda = \Lambda(\mathbb{G}) \subset T^*_\infty \mathbb{R}^2$ its associated Legendrian link, as defined in Section 2. See Section 2.3 for the definition of the sugar-free hull $\mathbb{S}_f$ of a face f in $\mathbb{G}$ and Section 4.8 for completeness. Note that the concept of sugar-free hulls, and whether a region is sugar-free, only depends on the behavior at nonconvex corners; see Definition 2.2.

Our main result, stated in Theorem 1.1, is the existence and explicit symplectic construction of a full quasicluster $\mathscr{A}$–structure on $\mathfrak{M}(\Lambda, T)$. In particular, the cluster $\mathscr{A}$–variables of the initial seed and all the once-mutated seeds are obtained by a new microlocal parallel transport along certain relative cycles on exact Lagrangian fillings of Λ. This microlocal parallel transport is associated to a sheaf quantization of each exact Lagrangian filling, following [Guillermou et al. 2012; Casals and Zaslow 2022], and we refer to it as a *microlocal merodromy*; see Section 4.

Theorem 1.1 (main result) *Let $\mathbb{G} \subset \mathbb{R}^2$ be a complete grid plabic graph, $\Lambda = \Lambda(\mathbb{G}) \subset (\mathbb{R}^3, \xi_{st})$ its associated Legendrian link, $T \subset \Lambda$ a set of marked points with at least one marked point per component of Λ, and $\mathfrak{M}(\Lambda, T)$ the stack of decorated microlocal rank-one constructible sheaves on $\mathbb{R}^2$ with singular support contained in Λ.*

Then there exists a canonical embedded exact Lagrangian filling $L = L(\mathbb{G}) \subset (\mathbb{R}^4, \omega_{st})$ of Λ and a canonical $\mathbb{L}$–compressing system $\mathfrak{S} = \{\gamma_1, \ldots, \gamma_\ell\}$ for L, indexed by the sugar-free hulls of $\mathbb{G}$, such that, for any completion of $\mathfrak{S}$ into a basis $\mathfrak{B}$ of $H_1(L, T)$, the following hold:

(i) *The microlocal merodromies A_{η_i}, defined on (and by using) the open chart $(\mathbb{C}^\times)^{b_1(L,T)} \subset \mathfrak{M}(\Lambda, T)$ associated to L, extend to global regular functions*

$$A_{\eta_i} \colon \mathfrak{M}(\Lambda, T) \to \mathbb{C}, \quad \text{ie } A_{\eta_i} \in \mathbb{O}(\mathfrak{M}(\Lambda, T)),$$

where $\mathfrak{B}^\vee = \{\eta_1, \ldots, \eta_s\}$ is the dual basis in $H_1(L \setminus T, \Lambda \setminus T)$.

(ii)   The microlocal merodromies $\{A_{\eta_1}, \ldots, A_{\eta_\ell}\}$ associated to the relative cycles that are dual to an $\mathbb{L}$–compressible absolute cycle in $\mathfrak{S}$ are irreducible functions in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$, whereas the merodromies $\{A_{\eta_{\ell+1}}, \ldots, A_{\eta_{b_1(L,T)}}\}$ are nonvanishing functions, ie units in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$.

(iii)   Let $L'_k \subset (\mathbb{R}^4, \omega_{\text{st}})$ be the Lagrangian filling obtained via Lagrangian disk surgery on L at the $\mathbb{L}$–compressing disk for $\gamma_k \in \mathfrak{S}$, and $\eta'_k \in H_1(L'_k \setminus T, \Lambda \setminus T)$ the image of η_k under the surgery. Then the merodromy $A_{\eta'_k}$ extends to a global regular function

$$A_{\eta'_k} : \mathfrak{M}(\Lambda, T) \to \mathbb{C}, \quad ie \ A_{\eta'_k} \in \mathbb{O}(\mathfrak{M}(\Lambda, T)),$$

and satisfies the cluster $\mathscr{A}$–mutation formula

$$A_{\eta'_k} A_{\eta_k} = \prod_{\eta_i \to \eta_k} A_{\eta_i} + \prod_{\eta_k \to \eta_j} A_{\eta_j}$$

with respect to the intersection quiver $Q(\mathfrak{B})$ of the basis elements $\mathfrak{B} \subset H_1(L, T)$.

Finally, the moduli variety $\mathfrak{M}(\Lambda, T)$ admits a cluster $\mathscr{A}$–structure with quiver $Q(\mathfrak{B})$ in the initial seed associated to the Lagrangian filling L, where the mutable vertices (dually) correspond to the absolute cycles in the $\mathbb{L}$–compressing system $\mathfrak{S}$ for L. Furthermore, different choices of completion of $\mathfrak{S}$ into a basis $\mathfrak{B}$ give rise to quasiequivalent cluster $\mathscr{A}$–structures.

The grid plabic graph $\mathbb{G}$ actually provides several natural completions of the $\mathbb{L}$–compressing system $\mathfrak{S}$ to a basis $\mathfrak{B}$, as explained in Section 3. The canonical exact Lagrangian filling $L = L(\mathbb{G})$ associated with $\mathbb{G}$ is obtained as the Lagrangian projection of the Legendrian surface whose front is given by the weave $\mathfrak{w}(\mathbb{G})$ associated with $\mathbb{G}$, which is constructed in Section 3. The weave $\mathfrak{w}(\mathbb{G})$ is used crucially in the argument so as to obtain a sheaf quantization of $L(\mathbb{G})$ and prove items (i)–(iii), as required. In addition to the existence of the cluster $\mathscr{A}$–structures on $\mathfrak{M}(\Lambda, T)$, another upshot of Theorem 1.1 is that the initial and the once-mutated cluster $\mathscr{A}$–variables can be named entirely in terms of symplectic topology, in an intrinsic and geometric manner. The resulting quasicluster $\mathscr{A}$–structure and these $\mathscr{A}$–variables can be equally considered and computed after a Hamiltonian isotopy.

In terms of the dichotomy between geometry and algebra, Theorem 1.1 shows that the ring $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ behaves *as if* it were always possible to perform an arbitrary sequence of Lagrangian disk surgeries starting at $L(\mathbb{G})$ with the curve configuration from the $\mathbb{L}$–compressing system $\mathfrak{S}$. It is known that geometric obstructions to further surger the Lagrangian skeleton can arise as one performs a series of Lagrangian surgeries (geometric mutations), eg through the appearance of immersed curves, or algebraic intersection numbers differing from geometric ones, and yet the existence of the cluster $\mathscr{A}$–structure built in Theorem 1.1 shows that it is not possible to detect such obstructions by studying $\mathbb{O}(\mathfrak{M}(\Lambda, T))$. Table 1 schematically relates different ingredients involved in the proof of Theorem 1.1.

There are several items from Theorem 1.1 that can be helpful to unpack. First, by a modification of the Guillermou–Jin–Treumann map — see [Jin and Treumann 2017] — the Lagrangian filling L yields

grid plabic graph $\mathbb{G}$	symplectic topology in $T^*\mathbb{R}^2$	cluster theory
alternating strand diagram (with marked points T)	Legendrian link $\Lambda \subseteq T^*_\infty \mathbb{R}^2$ (with marked points T)	D^-–stack $\mathfrak{M}(\Lambda, T)$ from dg category $\mathrm{Sh}_\Lambda(\mathbb{R}^2)$
Goncharov–Kenyon conjugate surface associated to $\mathbb{G}$	weave for Lagrangian filling L ($\implies$ sheaf quantization $\mathscr{F}(L)$)	open toric chart $T_L = (\mathbb{C}^\times)^{b_1(L,T)} \subseteq \mathfrak{M}(\Lambda, T)$
sugar-free hull of $\mathbb{G}$	$\mathbb{L}$–*compressible curve* $\gamma \subseteq L$ with dual relative cycle $[\eta] \in N = H_1(L \setminus T, \Lambda \setminus T)$	T_L–coordinate that extends to a *global regular* function $A_\eta : \mathfrak{M}(\Lambda, T) \to \mathbb{C}$
set S of sugar-free hulls	mutable sublattice $\mathbb{Z}^{|S|} \subseteq N$	mutable variables $\{A_\eta\}$ in T_L
non-sugar-free region of $\mathbb{G}$ (eg a non-sugar-free face)	immersed curve $\vartheta \subseteq L$ with dual relative cycle ϕ in N (ϑ represented by immersed Y–tree in weave)	T_L–coordinate extending to *nonvanishing* global regular function $A_\phi : \mathfrak{M}(\Lambda, T) \to \mathbb{C}$
subset of non-square-free regions chosen via Hasse diagram (different choices allowed)	sublattice $\mathbb{Z}^{b_1(L)-|S|} \subseteq N$ complement to sublattice $\mathbb{Z}^{|S|}$ (different complements)	frozen variables $\{A_\phi\}$ in T_L (quasicluster equivalent)
intersection form on absolute H_1 of conjugate surface	intersection form on $M = H_1(L, T)$ (and thus on dual $N = M^*$)	quiver $Q(\{A_\eta\}, \{A_\phi\})$ for T_L (different from naive $Q(\mathbb{G})$)
"mutation" at sugar-free hull (not necessarily a square face, result often not a plabic graph but represented by a weave)	*Lagrangian surgery* $L' = \mu_\gamma(L)$ and relative cycle $\eta' = \mu_\gamma(\eta)$ in L'; sheaf quantization $\mathscr{F}(L')$ via *weave mutation* at Y–tree for γ	$T_{L'}$–coordinate extending to a *global regular* function $A_{\eta'} : \mathfrak{M}(\Lambda, T) \to \mathbb{C}$ given by *cluster $\mathcal{A}$–mutation* at η

Table 1: Ingredients in the symplectic construction of upper cluster algebra for $\mathbb{O}(\mathfrak{M}(\Lambda, T))$.

an open toric chart $(\mathbb{C}^\times)^{b_1} \subset \mathfrak{M}(\Lambda, T)$, where $b_1 = \mathrm{rk}(H_1(L \setminus T, \Lambda \setminus T)) = \mathrm{rk}(H_1(L, T))$. The group $H^1(L; \mathbb{C}^\times) = \mathrm{Hom}(H_1(L; \mathbb{Z}), \mathrm{GL}_1(\mathbb{C}))$ accounts for the $\mathbb{C}^\times$–local systems on $L(\mathbb{G})$, and the modification accounts for the relative piece given by the marked points T; see Section 2.7 for details. By construction, microlocal merodromies are a priori functions on this particular chart $(\mathbb{C}^\times)^{b_1}$, and they visibly depend on L. In fact, in many cases they are (restrictions of) rational functions with nontrivial denominators and do not extend to global regular functions. Nevertheless, Theorem 1.1 shows that, remarkably, there is a particular set of such functions, indexed by a basis completion of the $\mathbb{L}$–compressing system $\mathfrak{S}$, whose elements extend to regular functions from $(\mathbb{C}^\times)^{b_1}$ to the entire moduli $\mathfrak{M}(\Lambda, T)$.

Second, the frozen cluster $\mathcal{A}$–variables in Theorem 1.1 have two geometric, markedly distinct, origins: absolute cycles in $H_1(L)$, and relative cycles with endpoints in T which are themselves not dual to any absolute cycle. The appearance of the former type of frozen variables, associated to absolute cycles, is an entirely new phenomenon, starting the study of $\mathbb{L}$–(in)compressible curves in Lagrangian fillings. (For example, we show that a Chekanov $m(5_2)$ already displays such features.) At least to date, all known instances of frozen variables of geometric origin were related to marked points, in line with the latter

type of frozens. The existence of a cluster structure on $\mathfrak{M}(\Lambda, T)$ with a particular quiver Q has neat applications to symplectic geometry, eg studying the possible relative Lagrangian skeleta containing L for the Weinstein relative pair $(\mathbb{C}^2, \Lambda)$; see below for more.

Third, item (iii) in Theorem 1.1 geometrically keeps track of certain relative cycles before and after a Lagrangian surgery: the data being analyzed is the change of a specific local system along that relative cycle (which itself changes topologically). This local system is obtained by applying the microlocal functor, with the target being the Kashiwara–Schapira stack $\mu\mathrm{Sh}_\Lambda$, to a sheaf quantization of L. In our proof of Theorem 1.1, the sheaf quantization is obtained thanks to the construction of the weave $\mathfrak{w}(\mathbb{G})$, which represents a (front of the) Legendrian lift of L. In fact, Section 3 will provide a diagrammatic method to draw those relative cycles before and after a weave mutation, and Section 4 provides a Lie-theoretic procedure to compute with such (microlocal) local systems. Note also that the geometric mutations are associated with sugar-free hulls, which are not necessarily square faces and might include lollipops; the fact that the calculus of weaves allows for these general mutations is crucial in order to conclude that the coordinate ring of $\mathfrak{M}(\Lambda, T)$ is an upper cluster algebra.

Finally, the symplectic geometry perspective naturally leads to a quasicluster $\mathscr{A}$–structure, rather than a cluster $\mathscr{A}$–structure. Indeed, the weave $\mathfrak{w}(\mathbb{G})$ canonically gives the $\mathbb{L}$–compressing system $\mathfrak{S}$, which yields a linearly independent subset of $H_1(L \setminus T, \Lambda \setminus T)$. Nonetheless, there are cases in which this subset does not span and a choice of basis completion is precisely what introduces the quasicluster ambiguity. In particular cases, such as $\mathbb{G}$ being a plabic fence, the $\mathbb{L}$–compressing system already gives a basis and hence $\mathfrak{M}(\Lambda(\mathbb{G}), T)$ carries a natural cluster $\mathscr{A}$–structure, but for a generic grid plabic graph $\mathbb{G}$ there is no a priori reason for that to be the case; the natural algebraic structure arising from symplectic geometry is only unique up to quasicluster equivalence.

Theorem 1.1 also implies a series of new computations and results in 3–dimensional contact topology. Indeed, in many interesting cases, such as those where the cluster algebra equals the upper cluster algebra [Muller 2013; 2014; 2022], the existence of a full cluster $\mathscr{A}$–structure on the moduli space[2] $\mathfrak{M} = \mathfrak{M}(\Lambda, T)$, as proven in Theorem 1.1, leads to:

(1) The computation of its de Rham cohomology ring $H^*(\mathfrak{M}, \mathbb{C})$, including the refinement of its mixed Hodge structure. These computations are done in [Lam and Speyer 2022] for the locally acyclic cases.

(2) The existence of a holomorphic (pre)symplectic structure for the moduli space $\mathfrak{M}$. This allows for many classical techniques, such as quantization, to be applied to the coordinate ring $\mathbb{O}(\mathfrak{M})$; see [Gekhtman et al. 2010]. We emphasize that the cluster $\mathscr{A}$–variables associated to a seed are exponential Darboux coordinates for the symplectic 2–form. Note also that a holomorphic symplectic structure on the augmentation variety was recently constructed in [Casals et al. 2020] by different means (using the Cartan 3–form and Bott–Shulman forms), and see work of P Boalch

---

[2]If not made explicitly, the set of marked points T is taken to have one marked point per component of Λ.

[2014a; 2014b]. Upcoming work with our collaborators will show that these holomorphic symplectic structures coincide whenever they can be compared.

(3)   In the Louise case [Lam and Speyer 2022], it is possible to compute the eigenvalues of the Frobenius automorphism on ℓ–adic cohomology and perform finite point counts $\#\mathfrak{M}(\mathbb{F}_q)$ over finite fields $\mathbb{F}_q$ for $q = p^k$ and p large enough. These ought to be compared with the contact and symplectic results in [Henry and Rutherford 2015; Ng et al. 2017].

Another byproduct of our result, thinking in terms of cluster ensembles [Fock and Goncharov 2006a], is that there also exists a (full) cluster $\mathscr{X}$–structure. Let $\mathcal{M}_1(\Lambda)$ be the undecorated stack associated to $\mathfrak{M}(\Lambda, T)$ and $\mathcal{M}_1(\Lambda, T)$ its enhancement with framing data at T. Theorem 1.1 implies the following result:

Corollary 1.2   *Let $\mathbb{G} \subset \mathbb{R}^2$ be a complete grid plabic graph, $\Lambda = \Lambda(\mathbb{G}) \subset (\mathbb{R}^3, \xi_{\text{st}})$ its associated Legendrian link and $T \subset \Lambda$ marked points. Then there exists a quasicluster $\mathscr{X}$–structure on $\mathcal{M}_1(\Lambda, T)$.*

In fact, each completion of the $\mathbb{L}$–compressing system $\mathfrak{S}$ to a basis $\mathfrak{B}$ of $H_1(L, T)$ gives a cluster $\mathscr{X}$–structure on $\mathcal{M}_1(\Lambda, T)$. The initial quiver Q is defined by the intersections in $\mathfrak{B}$ and the initial cluster $\mathscr{X}$–variables are microlocal monodromies associated with elements of $\mathfrak{B}$. In addition, the mutable cluster $\mathscr{X}$–variables are those associated with curves in the $\mathbb{L}$–compressing system $\mathfrak{S}$ and different choices of completion of $\mathfrak{S}$ to a basis $\mathfrak{B}$ give quasiequivalent cluster $\mathscr{X}$–structures on $\mathcal{M}_1(\Lambda, T)$.

Corollary 1.2 is a new result and establishes the existence of a (full) cluster $\mathscr{X}$–structure. It is crucial to understand that there is currently no proof of Corollary 1.2 on its own. Namely, we are only able to deduce the existence of a cluster $\mathscr{X}$–structure once we have proven the existence of a full cluster $\mathscr{A}$–structure in Theorem 1.1; two mathematical reasons are that the results used from [Berenstein et al. 2005] are only applicable to cluster $\mathscr{A}$–structures and that the codimension 2 arguments in Section 4 require an explicit understanding of the $\mathscr{A}$–variables, including their irreducibility; see also Section 2.8.

The two moduli $\mathfrak{M}(\Lambda(\mathbb{G}), T)$ and $\mathcal{M}_1(\Lambda(\mathbb{G}), T)$ in Theorem 1.1 and Corollary 1.2 form a cluster ensemble. In Section 5, we focus on shuffle grid plabic graphs and prove that these cluster varieties always admit a Donaldson–Thomas (DT) transformation. See [Kontsevich and Soibelman 2010; Goncharov and Shen 2018] for the necessary preliminaries on DT transformations. In fact, we realize this cluster automorphism geometrically, as a composition of a Legendrian isotopy of $\Lambda(\mathbb{G})$ and the strict contactomorphism $t : (x, y, z) \mapsto (-x, y, -z)$ of $(\mathbb{R}^3, \ker\{dz - y\,dx\})$. In particular, we conclude the following result:

Corollary 1.3   *Let $\mathbb{G}$ be a shuffle grid plabic graph. Consider the contactomorphism t and the half Kálmán loop Legendrian isotopy $K^{1/2}$. Then the composition $t \circ K^{1/2}$ induces the (unique) cluster Donaldson–Thomas transformation of $\mathcal{M}_1(\Lambda(\mathbb{G}))$.*

In particular, the cluster duality conjecture holds for the cluster ensemble $\big(\mathfrak{M}(\Lambda(\mathbb{G}), T), \mathcal{M}_1(\Lambda(\mathbb{G}), T)\big)$.

The explicit sequence of mutations realizing the DT transformation is presented in Section 5. We show it is a reddening sequence. Examples prove that it is not necessarily a maximal green sequence.

Finally, the contact and symplectic geometric results and techniques we use and develop to prove Theorem 1.1 are invariant under Hamiltonian isotopies, not necessarily compactly supported. Given that the cluster coordinates in Theorem 1.1 and Corollary 1.2 are all intrinsically named through symplectic geometric means; they can be named, and computed, after a compactly supported Hamiltonian isotopy is applied to $L(\mathbb{G})$ or a contact isotopy is applied to $\Lambda(\mathbb{G})$. This is a distinctive crucial feature which had been missing in [Casals et al. 2020; 2021; Casals and Zaslow 2022; Gao et al. 2020a], where even the initial seed could not typically be defined (nor computed) after a Legendrian isotopy.[3]

Notation   We denote by $[a, b]$ the discrete interval $\{k \in \mathbb{N} \mid a \le k \le b\}$ if $a \le b$ with $a, b \in \mathbb{N}$. In this article, S_n denotes the group of permutations of n elements for $n \in \mathbb{N}$, and s_i its ith simple transposition for $i \in [1, n-1]$. We abbreviate $s_{[b,a]} := s_b s_{b-1} \ldots s_{a+1} s_a$ and $s_{[b,a]}^{-1} := s_a s_{a+1} \ldots s_{b-1} s_b$ for $a < b$ with $a, b \in \mathbb{N}$, and $s_{[b,a]}$ and $s_{[b,a]}^{-1}$ are empty if $b < a$. Let $w_{0,n} \in S_n$ be the longest word in the symmetric group S_n; we will sometimes write $w_0 \in S_n$ if n is clear by context. The standard word $\mathtt{w}_{0,n}$ for $w_{0,n}$ is defined to be the reduced expression $\mathtt{w}_{0,n} := s_{[1,1]} s_{[2,1]} s_{[3,1]} \cdots s_{[n-1,1]}$.

Extended version   This paper is a condensed account of arXiv 2204.13244, also available on our research websites, which contains a series of additional examples and figures, as well as more detail in some of the proofs and motivation and context in parts of the construction. The interested reader might benefit from the more inviting extended version, as it is more comprehensive and builds the proofs in a more self-contained manner. That said, we believe experts will also appreciate this streamlined version, where only the logically necessary steps for our main results are included.

2   Grid plabic graphs and Legendrian links

In this section we introduce the starting characters in the manuscript. On the combinatorial side, we introduce the notion of a grid plabic graph $\mathbb{G}$ in Section 2.1, and that of sugar-free hulls in Section 2.3. On the geometric side, we introduce a front for the Legendrian link $\Lambda(\mathbb{G})$ associated to the alternating strand diagram of a grid plabic graph $\mathbb{G}$ in Section 2.4, and set up the necessary moduli spaces from the microlocal theory of sheaves in Section 2.7. Several explicit examples are provided in Section 2.5.

2.1   Grid plabic graphs

The input object in our results is the following type of graphs:

---

[3]The pullback structures from [Shende et al. 2019, Section 3] had the same issue.

Definition 2.1 An embedded planar bicolored graph $\mathbb{G} \subset \mathbb{R}^2$ is said to be a *grid plabic graph* (or *GP graph* for short) if it satisfies the following conditions:

(i) The vertices of $\mathbb{G} \subset \mathbb{R}^2$ belong to the standard integral lattice $\mathbb{Z}^2 \subset \mathbb{R}^2$, and they are colored in either black or white.

(ii) The edges of $\mathbb{G} \subset \mathbb{R}^2$ belong to the standard integral grid $(\mathbb{Z} \times \mathbb{R}) \cup (\mathbb{R} \times \mathbb{Z}) \subset \mathbb{R}^2$. Edges that are contained in $\mathbb{Z} \times \mathbb{R}$ are said to be *vertical*, and edges that are contained in $\mathbb{R} \times \mathbb{Z}$ are said to be *horizontal*.

(iii) A maximal connected union of horizontal edges is called a *horizontal line*. Each horizontal line must end at a univalent white vertex on the left and a univalent black vertex on the right. These univalent vertices are called *lollipops*.

(iv) Each vertical edge must end at trivalent vertices of opposite colors, and the endpoints of a vertical edge must be contained in the interior of a horizontal line.

In Definition 2.1, it is fine to allow for bivalent vertices. The Legendrian isotopy type of the zigzag diagram, as introduced in Section 2.4, does not change when inserting such vertices, nor does the Hamiltonian isotopy type of the Lagrangian filling associated to the conjugate surface.

2.2 Column types and associated transpositions

The intersection of a GP graph $\mathbb{G} \subset \mathbb{R}^2$ with a subset of the form $\{(x, y) \in \mathbb{R}^2 \mid l < x < r\} \subset \mathbb{R}^2$ for some $l, r \in \mathbb{R}$ with $l < r$ is said to be a column of $\mathbb{G}$. Any GP graph $\mathbb{G}$ is composed by the horizontal concatenation of three types of nonempty column, called *elementary columns*. These three types of elementary column are depicted in Figure 2 and can be described as follows:

- A column is said to be Type 1 if it solely consists of parallel horizontal lines, ie it contains no vertices.

- A column is said to be Type 2, or a crossing, if it contains exactly two oppositely colored vertices of $\mathbb{G}$ and a (unique) vertical edge between them.

- A column is said to be Type 3, or a lollipop, if it contains exactly one lollipop. Note that the lollipop can be either white or black.
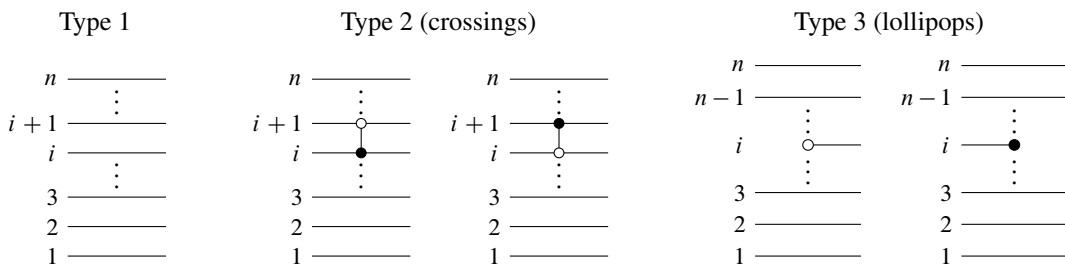
Figure 2: The three types of elementary column in a GP graph.

We label the horizontal $\mathbb{G}$–edges in Type 1 and 2 columns by consecutively increasing natural numbers from bottom to top. The horizontal lines of a Type 3 column are labeled in a similar way, but using the right side of the column in the case of a white lollipop and using the left side of the column in the case of a black lollipop. Without loss of generality, we always assume that there is a Type 1 column on each side of a column of Type 2 or 3.

Let $S_{\mathbb{N}}$ be the (infinite) group of permutations on the set $\mathbb{N}$. It is generated by simple transpositions $s_i = (i, i + 1)$ with $i \in \mathbb{N}$. Within $S_{\mathbb{N}}$, we define $S_{[a,b]} \cong S_{b-a+1}$ to be the subgroup consisting of bijections that map i back to itself for all $i \notin [a, b]$. As we scan from left to right across the elementary columns of $\mathbb{G}$, we associate a copy of $S_{[a,b]}$ for some $[a, b]$ with each column of Type 1 or 2 via these rules:

- We start with the empty set before the leftmost white lollipop, and we associate $S_{[1,1]}$ with the Type 1 column right after the leftmost white lollipop.

- The symmetric group $S_{[a,b]}$ does not change as we scan through a Type 1 or 2 column.

- If the symmetric group is $S_{[a,b]}$ before a Type 3 column with a white lollipop, then the symmetric group after this Type 3 column is $S_{[a,b+1]}$.

- If the symmetric group is $S_{[a,b]}$ before a Type 3 column with a black lollipop, then the symmetric group after this Type 3 column is $S_{[a+1,b]}$.

In summary, when passing through a white lollipop we move from a copy of S_k to a copy of S_{k+1} by adding a simple transposition at the end (with a larger subindex), and when passing through a black lollipop we move from a copy of S_{k+1} to a copy of S_k by dropping the first transposition (with smaller subindex).

2.3 Sugar-free hulls

By definition, a *face* of a GP graph $\mathbb{G}$ is any bounded connected component of $\mathbb{R}^2 \setminus \mathbb{G}$. A face is said to contain a lollipop if its closure in $\mathbb{R}^2$ contains a univalent vertex of $\mathbb{G}$. A *region* of a GP graph $\mathbb{G}$ is a union of faces whose closure in $\mathbb{R}^2$ is connected; in particular, a face is a region and the union of any pair of adjacent faces is a region. For instance, the yellow and red areas depicted in Figure 1 are both faces and the yellow face contains a lollipop; their union is a region (which will be the sugar-free hull of the yellow face).

The *boundary* ∂R of a region R is the topological (PL-smooth) boundary of its closure $\overline{R} \subset \mathbb{R}^2$. The boundary ∂R of a region necessarily consists of straight line segments meeting at corners that have either 90° or 270° angles. By definition, a 270° corner is said to be *left-pointing* if it is of the form $\ulcorner$ or $\llcorner$, and a 270° corner is said to be *right-pointing* if it is of the form $\urcorner$ or $\lrcorner$. Equipped with this terminology, we introduce the following notion:
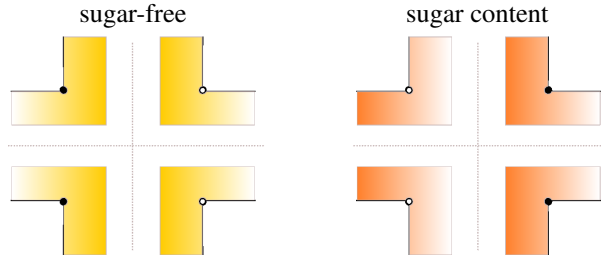
Figure 3: The four corners depicted on the left, in yellow, are allowed in a sugar-free region. The four corners depicted on the right, in orange, are not allowed in a sugar-free region; they have sugar content.

Definition 2.2 Given a grid plabic graph $\mathbb{G}$, a region R is said to be *sugar-free* if all left-pointing 270° corners along ∂R are white and all right-pointing 270° corners along ∂R are black. See Figure 3 for a picture with the allowed (and disallowed) corners. The *sugar-free hull* $\mathbb{S}(f)$ of a face f of a plabic graph $\mathbb{G}$ is defined to be the intersection of all sugar-free regions R containing f. In particular, sugar-free hulls are sugar-free regions.

The boundary of a sugar-free region has the following characterization, which follows immediately from the fact that all vertical bars must be of different colors at the two ends:

Lemma 2.3 *Let R be a sugar-free region in a GP graph $\mathbb{G}$. Then ∂R must be decomposed as a concatenation of staircases of the four types illustrated in Figure 4.*

Lemma 2.4 *Let $\mathbb{G}$ be a GP graph, $R \subset \mathbb{G}$ be a sugar-free region and C a column in $\mathbb{G}$ of any type. Then the intersection $R \cap C$ has at most one connected component.*

Proof By definition, the region R is connected. Thus, in order for the intersection $R \cap C$ to have more than one connected component, R needs to make a (horizontal) U-turn at some point and ∂R must contain a part that is of the shape "R (" or "$) R$", where the parentheses indicate the U-turn and the letter R indicates the side of the region. However, such a shape cannot be built using the four types of staircases in Lemma 2.3 and therefore $R \cap C$ can have at most one connected component. □

Figure 4: Four types of staircase building blocks for the boundary ∂R of a sugar-free region $R \subset \mathbb{G}$. In each instance, the letter R marks the location of the region in the plane. The dashed lines indicate that ∂R can continue in either of the two branches

Figure 5: The local models for an alternating strand diagram associated to a GP graph $\mathbb{G}$. The small hairs indicate the coorienting direction, which is needed to specify a Legendrian lift.

Note that if a region R is not simply connected, then there must exist a column C such that $R \cap C$ has more than one connected component. Thus, Lemma 2.4 has the following consequence, despite the fact that there may exist non–simply connected faces in the GP graph:

Corollary 2.5 *Sugar-free regions are simply connected.*

2.4 Legendrian links

In this subsection we introduce the Legendrian link $\Lambda(\mathbb{G}) \subset (\mathbb{R}^3, \xi_{\mathrm{st}})$ associated to a GP graph $\mathbb{G} \subset \mathbb{R}^2$ and explain how to algorithmically draw a specific front by scanning $\mathbb{G}$ left to right. Let us begin with the concise definition of $\Lambda(\mathbb{G})$:

Definition 2.6 Let $\mathbb{G} \subset \mathbb{R}^2$ be a GP graph. The Legendrian link $\Lambda(\mathbb{G}) \subset (\mathbb{R}^3, \xi_{\mathrm{st}})$ is the Legendrian lift of the alternating strand diagram of $\mathbb{G}$, understood as a cooriented front in $\mathbb{R}^2$, considered inside a Darboux ball in $(T_\infty^* \mathbb{R}^2, \xi_{\mathrm{st}})$.

Alternating strand diagrams were introduced in [Postnikov 2006, Definition 14.1] for a reduced plabic graph. In general, we associate such diagrams to a GP graph $\mathbb{G} \subset \mathbb{R}^2$ according to the two local models shown in Figure 5, where the hairs indicate the coorientation. The alternating strand diagram near a lollipop (or a bivalent vertex) is the same as in [Postnikov 2006], and the coorientation in these pieces is implied by the coorientations above.

By definition, the Legendrian lift of a cooriented immersed curve on the plane $\mathbb{R}^2$ is a Legendrian link inside the ideal contact boundary $(T_\infty^* \mathbb{R}^2, \xi_{\mathrm{st}})$. The contact structure is the kernel of the restriction of the Liouville 1–form on $T^* \mathbb{R}^2$ to this hypersurface. In general, such Legendrian links cannot be contained in a Darboux ball, but, for a GP graph $\mathbb{G}$, the Legendrian lift $\Lambda(\mathbb{G})$ is naturally contained in a Darboux ball, as we now explain. Let us choose Cartesian coordinates $(u, v) \in \mathbb{R}^2$. Then the contact structure on $T_\infty^* \mathbb{R}_{u,v}^2$ can be identified as the kernel of the contact 1–form $\alpha_{\mathrm{st}} := \cos\theta \, du + \sin\theta \, dv$, where $\theta \in [0, 2\pi)$ is the angle between a given covector $a \, du + b \, dv$ and du, $a, b \in \mathbb{R}$ and $a^2 + b^2 \neq 0$. Note that $T_\infty^* \mathbb{R}^2$ is diffeomorphic to $\mathbb{R}^2 \times S^1$ and $\theta \in S^1$ records that circle coordinate. In fact, we can consider the 1–jet space $(J^1 S^1, \xi_{\mathrm{st}})$ with its standard contact structure $\ker\{\beta_{\mathrm{st}}\}$, $\beta_{\mathrm{st}} := dz - y \, d\theta$, where $y \in \mathbb{R}$ is the coordinate along the cotangent fiber and $z \in \mathbb{R}$ the Reeb coordinate, as $J^1 S^1 := T^* S^1 \times \mathbb{R}$. Then there exists a strict contactomorphism $\varphi \colon (T^\infty \mathbb{R}^2, \alpha_{\mathrm{st}}) \to (J^1 S^1, \beta_{\mathrm{st}})$ given by

$$\varphi^*(\theta) = \theta, \quad \varphi^*(y) = -u\sin\theta + v\cos\theta, \quad \varphi^*(z) = u\cos\theta + v\sin\theta.$$
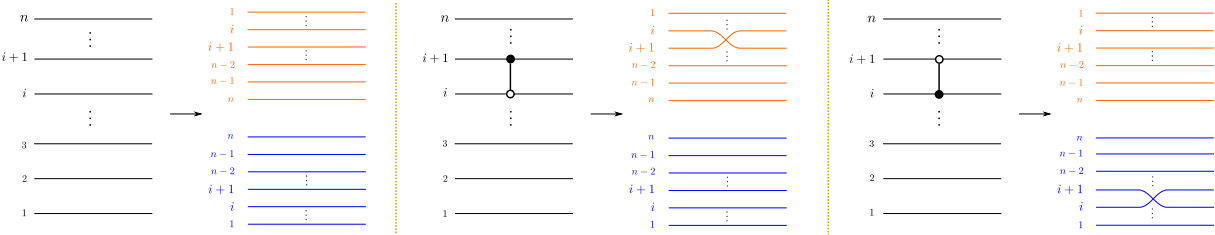
Figure 6: The rules to construct the front $\mathfrak{f}(\mathbb{G})$ from the elementary columns of a GP graph $\mathbb{G}$. In this case, Type 1 and Type 2 columns are depicted, with the GP graph $\mathbb{G}$ on the left and the front $\mathfrak{f}(\mathbb{G})$ on the right. We have colored the top n strands of the front in orange and the bottom n strands of the front in blue for clarification purposes.

For any open interval $I \subset S^1$, $(J^1 I, \xi_{st})$ is contactomorphic to a standard Darboux ball $(\mathbb{R}^3, \xi_{st})$. In consequence, if a cooriented immersed curve $\mathfrak{f} \subset \mathbb{R}^2$ in $\mathbb{R}^2$ has a Gauss map that misses one given angle θ_0, the Legendrian lift of $\mathfrak{f} \subset \mathbb{R}^2$ is contained in $(J^1(S^1 \setminus \theta_0), \xi_{st})$, which is contactomorphic to a Darboux ball. This happens for the alternating strand diagram of a GP graph $\mathbb{G} \subset \mathbb{R}^2$ and thus $\Lambda(\mathbb{G})$ naturally lives inside a Darboux ball.

Let us now construct a particular type of (wave)front for the Legendrian link $\Lambda(\mathbb{G})$, which is useful to describe our moduli spaces in Lie-theoretic terms. For that, we consider the front $\mathfrak{f}(\mathbb{G}) \subset \mathbb{R}^2$ obtained by dividing the GP graph $\mathbb{G}$ into elementary columns and then use the assignments depicted in Figures 6 and 7. Namely, to an elementary column of Type 1 with n strands, we assign a front consisting of $2n$ parallel horizontal strands. For an elementary column of Type 2 with n strands and a vertical bar at the ith position, we assign a front consisting of $2n$ parallel horizontal strands with a crossing at the ith position either at the top n strands or the bottom n strands, depending on whether the vertical bar has a white vertex at the top or at the bottom. Figure 6 depicts these three cases for Types 1 and 2. The case of an
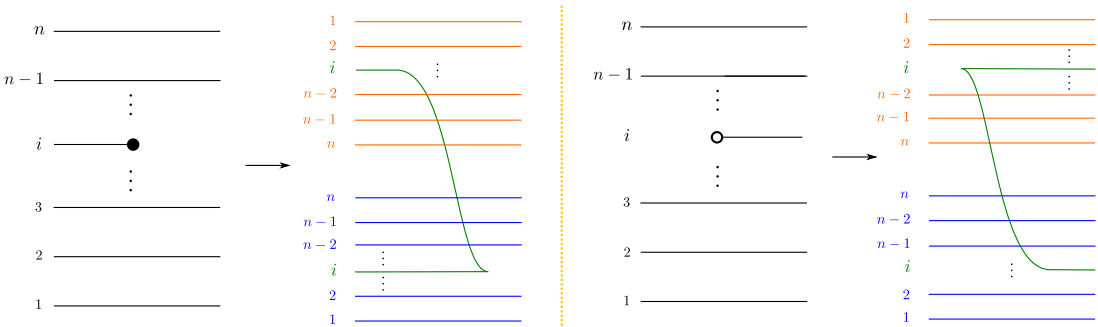
Figure 7: The rules to construct the front $\mathfrak{f}(\mathbb{G})$ from the elementary columns of a GP graph $\mathbb{G}$. In this case, the two kinds of Type 3 columns are depicted, with the GP graph $\mathbb{G}$ on the left and the front $\mathfrak{f}(\mathbb{G})$ on the right. We have colored the top n strands of the front in orange, the bottom n strands of the front in blue, and the newly inserted strand with a cusp in green, to help visualize the front.

elementary column of Type 3 involves inserting a right (resp. left) cusp at the ith position (plus some additional crossings) if there is a black (resp. white) lollipop inserted at the ith position. Figure 7 depicts the two possible cases for a Type 3 column.

Note that the $2n$ strands in the front are labeled in a specific manner in Figures 6 and 7, starting the count from the outer strand and increasing towards the middle. This choice of labeling is the appropriate one: in this way, when only left cusps have appeared, which is always the case at the beginning if we read $\mathbb{G}$ left to right, the ith top strand (in orange) and the ith bottom strand (in blue) coincide. Now, the front $\mathfrak{f}(\mathbb{G}) \subset \mathbb{R}^2$ lifts to a Legendrian link $\Lambda(\mathfrak{f}(\mathbb{G})) \subset (\mathbb{R}^3, \xi_{st})$. We observe that, in this case, the lift can be considered directly into $\mathbb{R}^3$, as the front is cooriented upwards and there are no vertical tangencies. The following proposition follows by applying the above contactomorphism $\varphi\colon (T^\infty \mathbb{R}^2, \alpha_{st}) \to (J^1 S^1, \beta_{st})$:

Proposition 2.7 *Let $\mathbb{G} \subset \mathbb{R}^2$ be a GP graph. Then the two Legendrian links $\Lambda(\mathbb{G}) \subset (\mathbb{R}^3, \xi_{st})$ and $\Lambda(\mathfrak{f}(\mathbb{G})) \subset (\mathbb{R}^3, \xi_{st})$ are Legendrian isotopic.*

2.5 Instances of GP graphs $\mathbb{G}$ and their Legendrian links $\Lambda(\mathbb{G})$

In this subsection we discuss a few examples of GP graphs $\mathbb{G}$ that lead to particularly interesting and well-studied Legendrian links.

Plabic fences Consider a GP graph $\mathbb{G} \subset \mathbb{R}^2$ whose white lollipops all belong to the line $\{-1\} \times \mathbb{R}$, and all black lollipops belong to the line $\{1\} \times \mathbb{R}$. Figure 8 depicts instances of such GP graphs. These GP graphs are called plabic fences in [Fomin et al. 2022, Section 12], following L Rudolph's fence terminology. It follows from Proposition 2.7 and the rules from Figures 6 and 7 that the Legendrian link $\Lambda(\mathbb{G})$ associated to a plabic fence $\mathbb{G} \subset \mathbb{R}^2$ is Legendrian isotopic to the (Legendrian lift of the) rainbow closure of a positive braid. In fact, given such a plabic fence $\mathbb{G} \subset \mathbb{R}^2$ with n horizontal lines, consider the positive braid word $\beta \in \mathrm{Br}_n^+$ whose k^{th} crossing is σ_j if and only if the kth vertical edge *with black on bottom* of $\mathbb{G}$ (starting from the left) is between the jth and $(j+1)^{st}$ horizontal strands. Similarly, consider the positive braid word δ whose m^{th} crossing is σ_{n-j} if and only if the mth vertical edge *with white on*
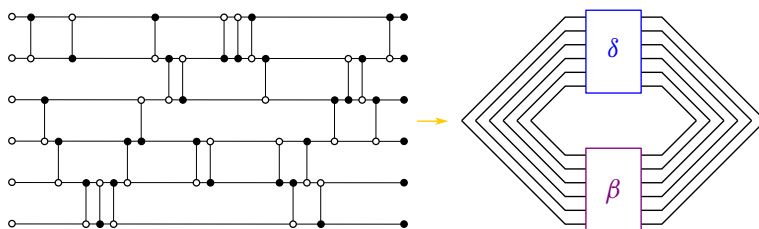
Figure 8: A front for the Legendrian link associated to the GP graph on the left is drawn on the right, where $\beta, \delta \in \mathrm{Br}_6^+$ are the positive braid words $\beta = \sigma_5 \sigma_1 \sigma_3 \sigma_4 \sigma_3 \sigma_5^2 \sigma_2 \sigma_1 \sigma_4$ and $\delta = \sigma_1 \sigma_3 \sigma_4 \sigma_5^2 \sigma_4 \sigma_1 \sigma_2 \sigma_4 \sigma_1 \sigma_2 \sigma_5 \sigma_4 \sigma_3 \sigma_2 \sigma_3 \sigma_1$.
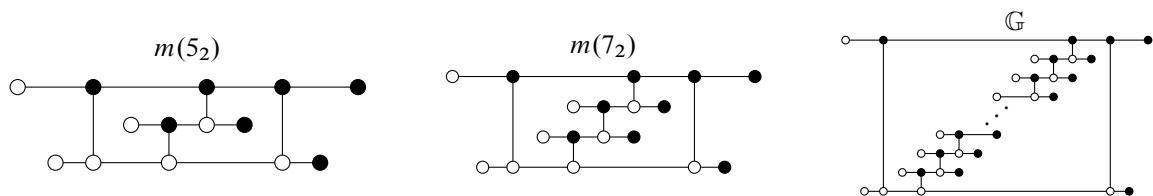
Figure 9: GP graphs whose alternating strand diagrams have the smooth type of (mirrors of) twist knots. The GP graph on the left (resp. middle) yields a max-tb Legendrian representative of $m(5_2)$ (resp. $m(7_2)$). The right illustrates the general case, where a GP graph is built by iteratively inserting — in a staircase manner — the local piece inside the GP graph in the upper left.

bottom of $\mathbb{G}$ is between the jth and $(j+1)^{\text{st}}$ horizontal strands. Then Figure 8, right, depicts a front for the Legendrian link $\Lambda(\mathbb{G})$, which is readily homotopic to the rainbow closure of the positive braid word $\beta\delta^\circ$ (or equivalently $\delta^\circ\beta$), where δ° denotes the reverse positive braid of δ.

Legendrian twist knots Let us consider the family of GP graphs $\mathbb{G}_n$, indexed by $n \in \mathbb{N}$, that we have depicted in Figure 9. Each GP graph $\mathbb{G}_n$ has two long horizontal bars and it is obtained by inserting a staircase with n steps between two vertical bars, themselves located at the leftmost and rightmost position. Figure 9 draws $\mathbb{G}_1$ (left) and $\mathbb{G}_2$ (middle). By using Figures 6 and 7, fronts for the associated Legendrian knots $\Lambda(\mathbb{G}_n)$ are readily drawn: Figure 10 depicts fronts for $\Lambda(\mathbb{G}_1)$ and $\Lambda(\mathbb{G}_2)$. In general, we conclude that $\Lambda(\mathbb{G}_n)$ is a max-tb Legendrian representative of a twist knot, with zero rotation number. Note that Legendrian twist knots are classified in [Etnyre et al. 2013]. In particular, the Legendrian knot $\Lambda(\mathbb{G}_1)$ associated to the GP graph depicted in Figure 9, left, is the unique max-tb Legendrian representative of $m(5_2)$ with a binary Maslov index. This is one half of the well-known Chekanov pair.[4]
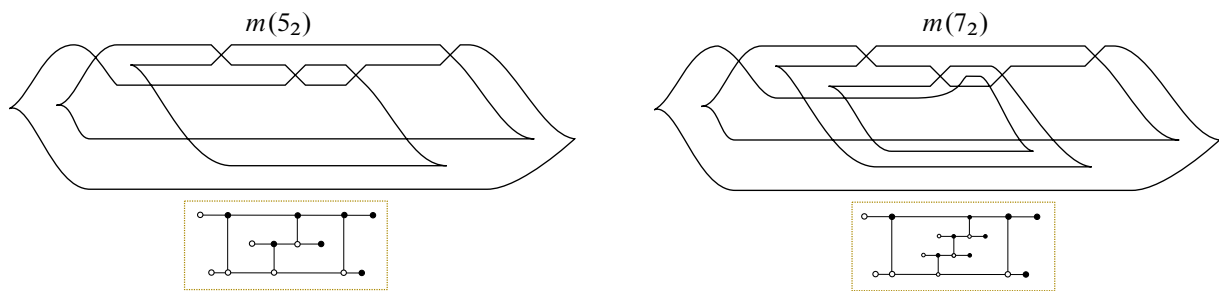
Figure 10: The two Legendrian fronts obtained from the left and middle GP graphs of Figure 9, according to our recipe translating from GP graphs to Legendrian front diagrams. The corresponding GP graphs are drawn in a small box below each front. The front diagram for the general case (Figure 9, right) is readily inferred from these two pictures: a knotted spiraling pattern is iteratively added to the center region of the front.

Shuffle graphs Let us introduce a class of GP graphs which leads to interesting examples.

---

[4]The other max-tb representative of $m(5_2)$ is not isotopic to $\Lambda(\mathbb{G})$ for any GP plabic graph $\mathbb{G}$.

Definition 2.8 A GP graph $\mathbb{G} \subset \mathbb{R}^2$ with n horizontal lines is said to be a *shuffle graph* if:

(1) There exist $M \in \mathbb{N}$ and $\sigma \in S_n$ such that each horizontal line goes from $(-M\sigma(i), i)$ to $(M\sigma(i), i)$.

(2) Vertical edges are all of the same pattern, ie they either all have a black vertex on top or they all have a white vertex on top.

The two families above, plabic fences and the GP graphs in Figure 9 for Legendrian twist knots, are instances of shuffle graphs. Shuffle graphs $\mathbb{G} \subset \mathbb{R}^2$ have the property that the Legendrian $\Lambda(\mathbb{G}) \subset (\mathbb{R}^3, \xi_{st})$ is Legendrian isotopic to the Legendrian lift of the (-1)–closure of a positive braid of the form $\beta\Delta$, where $\Delta \in \mathrm{Br}_n^+$ is the half-twist and $\beta \in \mathrm{Br}_n^+$ has Demazure product $\mathrm{Dem}(\beta) = w_0 = w_{0,n} \in S_n$. (By [Casals et al. 2020; Casals and Ng 2022], the condition $\mathrm{Dem}(\beta) = w_0 \in S_n$ is necessary.) For example, it is a simple exercise to verify that any (-1)–closure of a 3–stranded $\beta\Delta$, where $\beta, \Delta \in \mathrm{Br}_3^+$ and $\mathrm{Dem}(\beta) = w_0 \in S_3$, arises as $\Lambda(\mathbb{G})$ for some shuffle graph $\mathbb{G}$. In view of this and [Casals et al. 2020; 2021; Casals and Ng 2022], we refer to a positive braid $\beta \in \mathrm{Br}_n^+$ as Δ–*complete* if it is cyclically equivalent to a positive braid of the form $\Delta\gamma$, where Δ is the half twist and $\mathrm{Dem}(\gamma) = w_{0,n}$.

Let us point out two properties of the Legendrian links $\Lambda(\mathbb{G})$ that are useful. First, as we will explain, the Legendrian links $\Lambda(\mathbb{G})$ always bound an orientable exact embedded Lagrangian filling in the symplectization of $(\mathbb{R}^3, \xi_{st})$, and thus in the standard symplectic Darboux 4–ball. In particular, their Thurston–Bennequin invariant is always maximal and their rotation number vanishes. Second, it follows from the discussion in Section 2.4, especially Figures 6 and 7, that $\Lambda(\mathbb{G})$ admits a binary Maslov index and that the smooth type of $\Lambda(\mathbb{G})$ is that of the (-1)–closure of a positive braid. The former is particularly useful for us, as this implies that complexes of sheaves with singular support in $\Lambda(\mathbb{G})$ are quasi-isomorphic to sheaves (concentrated in degree 0) and it is thus possible to parametrize the moduli of objects of the appropriate dg category by an affine variety (or algebraic quotient thereof). Section 2.7 sets up the necessary ingredients on the microlocal theory of sheaves as it relates to these Legendrian links $\Lambda(\mathbb{G})$.

2.6 Lollipop chain reaction

In this subsection we introduce an algorithmic procedure, called a *lollipop chain reaction*, which aims to select faces for a sugar-free hull. The lollipop chain reaction initiates at a face f, and produces a collection of faces that are guaranteed to be inside the sugar-free hull $\mathbb{S}_f$. In many interesting cases of $\mathbb{G}$, such as shuffle graphs, this procedure yields the entire sugar-free hull $\mathbb{S}_f$. These combinatorial tools are used in Section 4.7, in the proof of Proposition 4.23. Let us start with the definition of a single *lollipop reaction*:

Definition 2.9 Let w be a white lollipop in a GP graph $\mathbb{G}$, and let h_1 and h_2 be the two adjacent horizontal $\mathbb{G}$–edges to the immediate left of w (in between which the lollipop appears). A vertical line segment between h_1 and h_2 is said to be a *wall*. By definition, the *lollipop reaction* initiated from the lollipop w pushes this wall to the right along $\mathbb{G}$ with the following rules: the wall shrinks or expands

according to the following five pictures and otherwise the wall stays between the same $\mathbb{G}$–edges:

| wall starts | wall shrinks | wall shrinks | wall expands | wall expands |
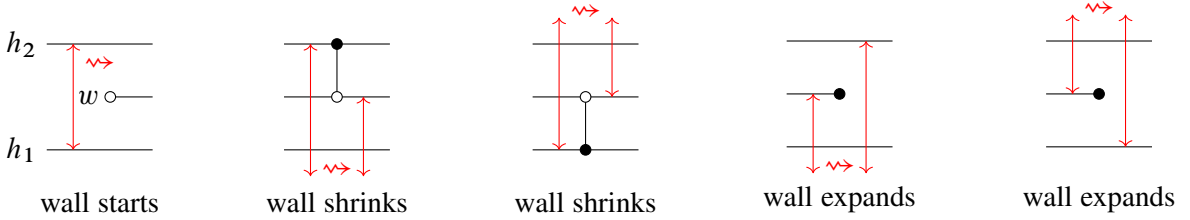
A single lollipop reaction initiated from a black lollipop b is defined in a symmetric fashion: start with a wall going between the two adjacent horizontal lines to the right of b, consider a wall between them and scan to the left. For a black lollipop, the wall shrinks or expands as it moves left according to the following five pictures and otherwise stays between the same $\mathbb{G}$–edges:

| wall starts | wall shrinks | wall shrinks | wall expands | wall expands |
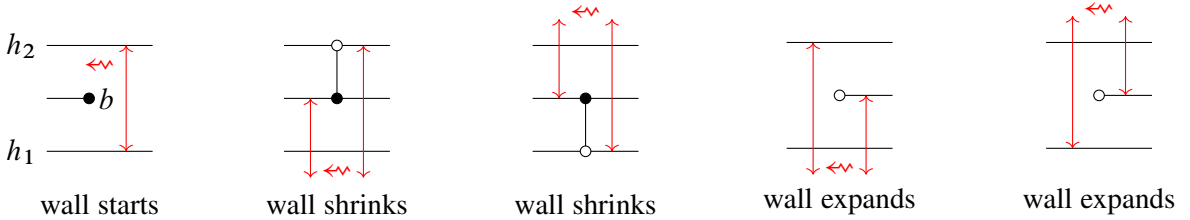
As the wall moves to the right (for a white lollipop) or to the left (for a black lollipop), we select all the faces that this wall scans through. By definition, a lollipop reaction *completes* when the length of the wall becomes zero. The output of a lollipop reaction is the selection of faces of the GP graph which it has scanned through. If the length of the wall becomes infinite (ie going to the unbounded region), then the lollipop reaction is said to be *incomplete*, and it outputs nothing.

In order to be effective, these lollipop reactions in general need to be iterated as follows:

Definition 2.10 Let $f \subset \mathbb{G}$ be a face of a GP graph $\mathbb{G}$. A *lollipop chain reaction* initiated at f is the recursive face selection procedure obtained as follows. First, select the face f. Then, for each of the newly selected faces and each inward-pointing lollipop of this face, run a single lollipop reaction and select new faces (if any).

Since the number of faces in $\mathbb{G}$ is finite, this process terminates either when no new faces are selected, for which we say that the lollipop chain reaction is *complete*, or when one of the single chain reactions is incomplete, for which we say that the whole lollipop chain reaction is *incomplete*.

Note that a single lollipop reaction selects faces that are *minimally needed* to avoid sugar-content corners on the immediate right of a white lollipop or the immediate left of a black lollipop. Therefore, the outcome of the lollipop chain reaction initiated from a face f must be contained in the sugar-free hull $\mathbb{S}_f$. In other words, if the lollipop chain reaction initiated from f is incomplete, then $\mathbb{S}_f$ does not exist. On the other hand, when sugar-free hulls $\mathbb{S}_f$ exist, lollipop chain reactions do produce sugar-free hulls for a large family of GP graphs:

Proposition 2.11 *Let $\mathbb{G}$ be a shuffle graph and f a face of $\mathbb{G}$ for which $\mathbb{S}_f$ is nonempty. Then the lollipop chain reaction initiated from f is complete and $\mathbb{S}_f$ coincides with the outcome of this lollipop chain reaction.*

Proof We observe that, in a shuffle graph there cannot be any black lollipop on the left side of a white lollipop, nor can there be any white lollipop on the right side of a black lollipop. Therefore, if there is a white lollipop inside a face f, then the part of the boundary ∂f straightly left of the white lollipop can only consists of a single vertical bar. Similarly, if there is a black lollipop inside a face f, then the part of the boundary ∂f straightly right of the black lollipop can only consists of a single vertical bar as well. Now, if the face f does not have any lollipops, then $\mathbb{S}_f = f$, which is also equal to the outcome of the lollipop chain reaction, as required. It remains to consider faces that do have lollipops inside. Without loss of generality, let us suppose that only vertical $\mathbb{G}$–edges of the type ⚬̣ appear in $\mathbb{G}$ and suppose that f contains a white lollipop. Consider the leftmost white lollipop w of f. Then, at the starting point, the wall for this lollipop goes between two adjacent horizontal lines h_1 and h_2, and ∂f only has a single vertical bar to the left of this wall. Then the lollipop reaction starts moving the wall to the right, and one of the following two situations must occur:

- *The wall never expands.* If this is the case, the wall must be shrinking towards the top as shown below:
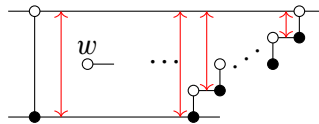
The result of the lollipop reaction is sugar-free.

- *The wall expands at some point.* Note that the wall only expands when it passes through a black lollipop b. Let g be the face containing b. Then the part of ∂g straightly right of b only consists of a single vertical edge e, and hence the rightward scanning must end at e. Note that in this case there can be a concavity below the selected faces right before the expansion of the wall. So we have

or
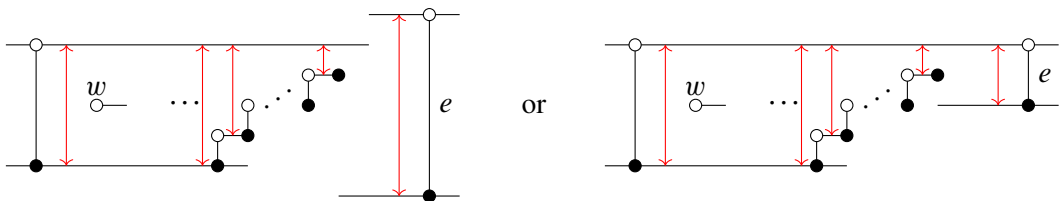
Now we start scanning leftward from the rightmost black lollipop in the face g. Note that, for a leftward scanning, the wall should be shrinking towards the bottom. Also, due to the "last in, first out" order on the horizontal lines, the top vertex of the vertical edge e cannot be below the top horizontal line of the previous scanning. If the top vertex of e is above the previous horizontal line, then the bottom vertex of e must be below the previous bottom horizontal line, and the leftward scanning will not stop until it goes back all the way to the beginning point of the previous scanning. On the other hand, if the top vertex

of e is on the previous horizontal line, then the bottom vertex of e can be above the previous bottom horizontal line. But then the horizontal line at the bottom of the vertical edge e must extend to the left and meets the staircase of the previous rightward scanning, and that is where the leftward scanning stops. In consequence, the lollipop reaction from the rightmost black lollipop of the face g must fill in the lower concavity of the previous rightward scanning.

Note that, in the second case above, the leftward scanning can also end in two ways, but we can conclude by induction that in the end all concavities will be filled and hence the resulting union must be sugar-free, as required.   □

There are many nonshuffle GP graphs for which lollipop chain reactions also yield sugar-free hulls, and thus the hypothesis in Proposition 2.11 is sufficient but not necessary.

2.7 Legendrian invariants from the microlocal theory of sheaves

In this subsection we lay out the necessary ingredients of the microlocal theory of constructible sheaves that we shall use in our contact-geometric framework. We describe the general setup in Section 2.7.1, based on [Kashiwara and Schapira 1985; 1990; Guillermou et al. 2012; Guillermou and Schapira 2014; Shende et al. 2017].[5] Section 2.7.2 discusses the specific simplifications that occur for the Legendrian links $\Lambda(\mathbb{G})$ and Section 2.7.3 introduces the necessary decorated version of the moduli stacks being discussed. Let $\mathbb{k}$ be a commutative coefficient ring, for us either $\mathbb{k} = \mathbb{Z}$ or $\mathbb{k} = \mathbb{C}$. Consider a smooth manifold M, $\pi_M : T^*M \to M$ its cotangent bundle and $T^*_\infty M \to M$ its ideal contact boundary; we will only need $M = \mathbb{R}^2$ and $\mathbb{R}^3$.

2.7.1 The general setup The general results on the microlocal theory of constructible sheaves were pioneered by M Kashiwara and P Schapira [1990] and, more recently, in collaboration with S Guillermou in [Guillermou et al. 2012]. The first category that we need is defined as follows:

Definition 2.12 The category $\mathbb{I}(\mathbb{k}_M)$ is the full dg subcategory of the dg category of locally bounded complexes of sheaves of $\mathbb{k}$–modules on M which consist of h–injective complexes of injective sheaves. The homotopy category of $\mathbb{I}(\mathbb{k}_M)$ is denoted by $[\mathbb{I}(\mathbb{k}_M)]$.

The dg category $\mathbb{I}(\mathbb{k}_M)$ is a strongly pretriangulated dg category and the six-functor formalism lifts to this dg enhancement $\mathbb{I}(\mathbb{k}_M)$; see [Schnürer 2018]. The homotopy category $[\mathbb{I}(\mathbb{k}_M)]$ is triangulated equivalent to the locally bounded derived category of sheaves on M, often denoted by $D^{\mathrm{lb}}(\mathbb{k}_M)$. For an object $F \in \mathbb{I}(\mathbb{k}_M)$, we denote by $\mu\mathrm{supp}(F) \subset T^*M$ its singular support understood as an object in $[\mathbb{I}(\mathbb{k}_M)] \simeq D^{\mathrm{lb}}(\mathbb{k}_M)$. The notion of singular support leads to defining the following dg categories:

---

[5]See also Guillermou's notes for his lecture series at the conference *Symplectic topology, sheaves and mirror symmetry* at the IMJ-PRG in Paris (2016) and [Shende et al. 2019].

Definition 2.13 Let $S \subset T^*M$ be a subset. The category $\mathbb{I}_S(\Bbbk_M)$ is the subcategory of $\mathbb{I}(\Bbbk_M)$ consisting of objects $F \in \mathbb{I}(\Bbbk_M)$ such that $\mu\text{supp}(F) \subset S$. The category $\mathbb{I}_{(S)}(\Bbbk_M)$ is the subcategory $\mathbb{I}(\Bbbk_M)$ consisting of objects $F \in \mathbb{I}(\Bbbk_M)$ for which there exists an open neighborhood Ω such that $\mu\text{supp}(F) \cap \Omega \subset S$.

Let $\Lambda \subset T^*_\infty M$ be a Legendrian submanifold. We denote by $\mathbb{I}_\Lambda(\Bbbk_M)$ and $\mathbb{I}_{(\Lambda)}(\Bbbk_M)$ the categories as above with the choice of subset S being the Lagrangian cone of Λ union the zero section $M \subset T^*M$.

The assignment $U \mapsto \mathbb{I}(\Bbbk_U)$ to each open subset $U \subset M$ is a stack of dg categories. Similarly, the prestack $\mathbb{I}_{(\Lambda)}$ defined by

$$\mathbb{I}_{(\Lambda)}(U) := \mathbb{I}_{(T^*U \cap \Lambda)}(\Bbbk_U), \quad U \subset M \text{ open},$$

is a stack. This is an advantage of using the injective dg enhancements instead of derived categories. A central result in symplectic topology [Guillermou et al. 2012] is that the stack $\mathbb{I}_{(\Lambda)}$ on M is a Legendrian isotopy invariant of the Legendrian $\Lambda \subset T^*_\infty M$. There are two constructions associated to the stack $\mathbb{I}_{(\Lambda)}$, as follows:

(i) **The microlocal functor $\mathfrak{m}_\Lambda$** The Kashiwara–Schapira stack $\mu\text{Sh}(\Bbbk_\Lambda)$ is the stack on Λ associated to the prestack

$$V \mapsto \mathbb{I}_{(V)}(\Bbbk_M; V), \quad V \subset \Lambda \text{ open},$$

where $\mathbb{I}_{(V)}(\Bbbk_M; V)$ is the Drinfeld dg quotient of $\mathbb{I}_{(V)}(\Bbbk_M)$ by $\mathbb{I}_{T^*M \setminus (V \cup M)}(\Bbbk_M)$. See [Kashiwara and Schapira 1990; Guillermou et al. 2012]. The quotient functor gives a functor of stacks

$$\mathfrak{m}_\Lambda : \mathbb{I}_{(\Lambda)} \to (\pi_M|_\Lambda)_*(\mu\text{Sh}(\Bbbk_\Lambda)).$$

Our use of this functor is twofold: in the case that $\Lambda \subset T^*_\infty \mathbb{R}^2$ is a Legendrian link, and in the case where $\Lambda \subset T^*_\infty \mathbb{R}^3$ is a Legendrian surface obtained as the lift of an exact Lagrangian filling of a Legendrian link.

(ii) **The moduli stack $\mathcal{M}_{\mathbb{I}_{(\Lambda)}(M)}$** By [Nadler 2016, Theorem 3.21], the global sections $\mathbb{I}_{(\Lambda)}(M)$ is a dg category equivalent to the category of (pseudo)perfect modules of a finite-type category, namely of the category of wrapped constructible sheaves $\text{Sh}^w_\Lambda(M)$ defined in [Nadler 2016, Definition 3.17]. Then the main result of [Toën and Vaquié 2007] implies that there exists a locally geometric D^-– stack $\mathcal{M}_{\mathbb{I}_{(\Lambda)}(M)}$, locally of finite presentation, which acts as the moduli stack of objects in the dg category $\mathbb{I}_{(\Lambda)}(M)$.

Finally, as explained in [Jin and Treumann 2017, Section 1.7], given an embedded exact Lagrangian filling $L \subset T^*M$ of a Legendrian submanifold $\Lambda \subset T^*_\infty M$, the microlocal functor $\mathfrak{m}_{\overline{L}}$ applied to the Legendrian lift $\overline{L} \subset J^1(M)$ yields an equivalence of categories between (a subcategory of) $\mathbb{I}_{(\overline{L})}(M)$ and the dg derived category of local systems on L. This induces an open inclusion $\iota_L : \mathbb{R}\text{Loc}(L) \to \mathcal{M}_{\mathbb{I}_{(\Lambda)}(M)}$, where $\mathbb{R}\text{Loc}(L)$ denotes the derived moduli space of local systems on L.

2.7.2 The concrete models For a Legendrian link $\Lambda \subset T_\infty^* \mathbb{R}^2$ with vanishing rotation number, the category of global sections of the Kashiwara–Schapira stack $\mu\mathrm{Sh}(k_\Lambda)$ admits a simple object Ξ by [Guillermou 2023, Part 10]. In addition, the functor $\mu\hom(\Xi, \cdot)$ is an explicit equivalence between $\mu\mathrm{Sh}(k_\Lambda)$ and the (twisted) stack Loc_Λ of (twisted) local systems on Λ.[6] In consequence, the microlocal functor $\mathfrak{m}_\Lambda$ described above yields a functor

$$\mathfrak{m}_{\Lambda, \Xi} \colon \underline{\mathbb{I}}_{(\Lambda)}(M) \to \mathrm{Loc}_\Lambda(\Lambda),$$

where we have considered global sections and identified the codomain of $\mathfrak{m}_\Lambda$ with Loc_Λ via $\mu\hom(\Xi, \cdot)$ and a choice of spin structure. In addition, given a Legendrian link $\Lambda \subset T_\infty^* \mathbb{R}^2$, we only need to consider the moduli substack $\mathcal{M}_1(\Lambda)$ of $\mathcal{M}_{\underline{\mathbb{I}}_{(\Lambda)}(\mathbb{R}^2)}$ which is associated to the subcategory of objects in $\underline{\mathbb{I}}_{(\Lambda)}(\mathbb{R}^2)$ whose image under $\mathfrak{m}_\Lambda$ is a local system (on Λ) of locally free $\Bbbk$–modules of rank one supported in degree zero. In the case that Λ admits a binary Maslov index, the stack $\mathcal{M}_1(\Lambda)$ is equal to its truncation $t_0(\mathcal{M}_1(\Lambda))$, which is an Artin stack.

Now, given an embedded exact Lagrangian filling $L \subset T^*M$ of Λ, the derived stack Loc_L of local systems on L is also equivalent to its truncation and the open inclusion ι_L described gives an inclusion $\iota_L \colon \mathrm{Loc}_1(L) \to \mathcal{M}_1(\Lambda)$ of Artin stacks, where $\mathrm{Loc}_1(L)$ are local systems (on L) of locally free $\Bbbk$–modules of rank one supported in degree zero. Since abelian local systems on L can be parametrized by $H^1(L, \Bbbk^\times)$, the inclusion ι_L provides a toric chart $\iota_L(\mathrm{Loc}_1(L))$ in the moduli stack $\mathcal{M}_1(\Lambda)$. In this article, we typically consider the ground ring $\Bbbk = \mathbb{C}$. If we are given a Legendrian link for which the stabilizers of $\mathcal{M}_1(\Lambda)$ are trivial and $\mathcal{M}_1(\Lambda)$ is smooth, then $\mathcal{M}_1(\Lambda)$ is (represented by) a smooth affine variety and an embedded exact Lagrangian filling L of Λ yields a toric chart $\iota_L \colon (\mathbb{C}^\times)^{2g(L)} \to \mathcal{M}_1(\Lambda)$, where $g(L)$ is the topological genus of the surface L.

Finally, both the inclusions $\iota_L \colon \mathrm{Loc}_1(L) \to \mathcal{M}_1(\Lambda)$ and the microlocal functors $\mathfrak{m}_\Lambda \colon \mathcal{M}_1(\Lambda) \to \mathrm{Loc}_1(\Lambda)$ can be computed explicitly from the front via cones of maps between stalks (of the sheaves parametrized by $\mathcal{M}_1(\Lambda)$). Indeed, we shall use the combinatorial model in [Shende et al. 2017, Section 3.3], where the points of $\mathcal{M}_1(\Lambda)$ are parametrized by functors from the poset category associated to the stratification induced by the Legendrian front to the abelian category of $\Bbbk$–modules (modulo acyclic complexes). In the case of Legendrian weaves, this combinatorial model is explained in [Casals and Zaslow 2022, Section 5].

2.7.3 A decorated moduli space Let $T = \{t_1, \dots, t_s\}$ with $t_i \subset \Lambda$ for $i \in [1, s]$ be a set of distinct points in a Legendrian link $\Lambda \subset T_\infty^* \mathbb{R}^2$. The elements of T will be referred to as *marked points*. The moduli stack $\mathcal{M}_1(\Lambda)$ can be decorated with additional trivializing information once a set of marked points T has been fixed, as follows:

Definition 2.14 Let $\Lambda \subset T_\infty^* \mathbb{R}^2$ be a Legendrian link with a fixed choice of Maslov potential and spin structure. Consider a set of marked points $T = \{t_1, \dots, t_s\}$ and label the components of $\Lambda \setminus T$ by Λ_i for

---

[6]A choice of spin structure on Λ and corresponding choices of spin structures for the Lagrangian fillings we consider allow a further identification to actual (untwisted) local systems. We implicitly have these choices in the background and translate them combinatorially in Section 4, through sign curves, when they are needed to assign signs.

$i \in \pi_0(\Lambda \setminus T)$. The moduli stack $\mathfrak{M}(\Lambda, T)$ is

$$\mathfrak{M}(\Lambda, T) := \{(F; \phi_1, \ldots, \phi_{|\pi_0(\Lambda \setminus T)|}) \mid F \in \mathcal{M}_1(\Lambda), \phi_i \text{ trivialization of } \mathfrak{m}_\Lambda(F) \text{ on } \Lambda_i\}.$$

Note that an abelian local system can always be trivialized over Λ_i if $\Bbbk = \mathbb{C}^*$. For a general ground ring $\Bbbk$, we require that there exist at least one marked point per component of Λ.

In Definition 2.14, the identification of (global sections of) the codomain of $\mathfrak{m}_\Lambda$ with the stack of local systems is fixed by the choice of Maslov potential and spin structure on Λ. There are at least two advantages to decorating the moduli stack of sheaves $\mathcal{M}_1(\Lambda)$ to $\mathfrak{M}(\Lambda, T)$. First, introducing the data of the trivializations in $\mathfrak{M}(\Lambda, T)$ often results in a smooth affine variety, even if $\mathcal{M}_1(\Lambda)$ was singular; this is similar to the classical setup with character varieties [Fock and Goncharov 2006b]. Second, the trivializations in $\mathfrak{M}(\Lambda, T)$ can be used to define global regular functions. In fact, we will show that $\mathfrak{M}(\Lambda, T)$ admits a cluster $\mathcal{A}$–structure, and our construction of the cluster $\mathcal{A}$–variables crucially relies on the existence of these decorations. Finally, the moduli space $\mathcal{M}_1(\Lambda, T)$ is defined similarly, by considering sheaves in $\mathcal{M}_1(\Lambda)$ with the additional data of trivializations of the stalks of the associated microlocal local systems at each of the marked points in T.

2.8 A clarification on the notion of cluster structures

In the literature, the sentence "a space Y has a cluster structure" has different meanings. We record here the precise definitions that have been used, implicitly or explicitly, and clarify the type of results we obtain. Let Q be a quiver, or more generally a skew-symmetrizable matrix. Consider the following concepts:

- The cluster algebra $\mathbb{A}_Q$. This is a commutative algebra and it comes endowed with a (typically infinite) system of generators $A_i \in \mathbb{A}_Q$, called the *cluster variables*. The vertices of the quiver give some of these cluster variables, and the other cluster variables are produced by the process of mutation. Cluster algebras were first introduced and studied by Fomin and Zelvinsky [1999; 2002; 2003]. The affine scheme associated to $\mathbb{A}_Q$ is $\mathrm{Spec}(\mathbb{A}_Q)$.

- The space $\mathcal{A}_Q$, called the cluster $\mathcal{A}$–space or cluster K_2–space, is a scheme obtained by birationally gluing certain tori according to Q. The ring of regular functions $\mathbb{O}(\mathcal{A}_Q) = \Gamma(\mathcal{A}_Q, \mathbb{O}_{\mathcal{A}_Q})$ is often referred to as the upper cluster algebra, due to its connection to [Berenstein et al. 2005]. This scheme is typically not finitely generated, but it is separated by [Gross et al. 2015, Theorem 3.14].

- The space $\mathcal{X}_Q$, called the cluster $\mathcal{X}$–space or cluster Poisson space, is also a scheme obtained by birationally gluing certain tori according to Q; the gluing maps are different than for $\mathcal{A}_Q$ above. The ring of regular functions $\mathbb{O}(\mathcal{X}_Q) = \Gamma(\mathcal{X}_Q, \mathbb{O}_{\mathcal{X}_Q})$ does not have a name. This scheme is typically not separated; see [Gross et al. 2015, Remark 2.6].

- The subset $\mu^{\mathcal{A}}_{\leq 1}(Q) \subset \mathcal{A}_Q$ consisting of the union of $\mathcal{A}$–tori associated to the initial seed for Q and its adjacent seeds, ie those obtained by performing *one* cluster $\mathcal{A}$–mutation. The ring of functions $\mathbb{O}(\mu^{\mathcal{A}}_{\leq 1}(Q))$ is known as the upper bound.

- The subset $\mu^{\mathcal{X}}_{\leq 1}(Q) \subset \mathcal{X}_Q$ consisting of the union of $\mathcal{X}$–tori associated to the initial seed for Q and its adjacent seeds, ie those obtained by performing *one* cluster $\mathcal{X}$–mutation.

The $\mathcal{A}$ and $\mathcal{X}$–schemes were first introduced and studied by Fock and Goncharov [2006b; 2006a] and subsequently featured in [Gross et al. 2015; 2018]. There is also the notion of a partial $\mathcal{X}$–structure (and partial $\mathcal{A}$–structure), as introduced in [Shende et al. 2019, Definition 5.11], where only some tori in $\mu^{\mathcal{X}}_{\leq 1}(Q)$ are considered. After studying the literature and discussing with experts, our conclusion is that "a space Y has a cluster structure" might mean that Y is equal to either of the (often quite different) spaces above, or even that it is equal up to codimension 2, ie $\mathbb{O}(Y)$ equals any of the (often quite different) rings of functions above. In certain cases, such as [Goncharov and Kontsevich 2021], it might also mean having a partial $\mathcal{A}$– or partial $\mathcal{X}$–structure for what the authors referred to as a noncommutative stack.

Remark 2.15 It is crucial to have a rigorous definition of the "space" Y and its "ring of functions" $\mathbb{O}(Y)$ so as to give precise meaning to the notion of admitting a cluster structure. If Y is a scheme, the sheaf of regular functions is well understood [Hartshorne 1977]. In our case, $\mathfrak{M}(\Lambda, T)$ are always affine schemes and $\mathcal{M}_1(\Lambda)$ and $\mathcal{M}_1(\Lambda, T)$ are always algebraic quotients of affine schemes.

Now, the spaces $\mathrm{Spec}(\mathbb{A}_Q)$, $\mathcal{A}_Q$, $\mathcal{X}_Q$, $\mu^{\mathcal{A}}_{\leq 1}(Q)$ and $\mu^{\mathcal{X}}_{\leq 1}(Q)$ are often quite different from each other, but the following facts hold:

(1) The inclusion $\mathbb{A}_Q \subset \mathbb{O}(\mathcal{A}_Q)$ always holds. This is a nontrivial fact known as the Laurent phenomenon. In particular, *all* cluster $\mathcal{A}$–variables $A_i \in \mathbb{A}_Q$ belong to $\mathbb{O}(\mathcal{A}_Q)$. In fact, $\mathbb{A}_Q$ can be defined to be the subalgebra of $\mathbb{O}(\mathcal{A}_Q)$ generated by the cluster $\mathcal{A}$–variables. In stark contrast, the cluster $\mathcal{X}$–variables X_i are almost never elements of $\mathbb{O}(\mathcal{X}_Q)$.

(2) The inclusion $\mathbb{A}_Q \subset \mathbb{O}(\mathcal{A}_Q)$ of the cluster algebra into its upper cluster algebra may or may not be an equality. This is referred to as the $\mathcal{A} = \mathcal{U}$ problem; see eg [Berenstein et al. 2005; Muller 2014]. The inclusion $\mathbb{O}(\mathcal{A}_Q) \subset \mathbb{O}(\mu^{\mathcal{A}}_{\leq 1}(Q))$ of the upper cluster algebra into its upper bound may or may not be an equality. It is known to be an equality for the case of full rank. In general, the equality $\mathbb{O}(\mathcal{X}_Q) = \mathbb{O}(\mu^{\mathcal{X}}_{\leq 1}(Q))$ always holds.

The main spaces Y we study here are the affine schemes $\mathfrak{M}(\Lambda, T)$. The results we prove imply that $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ equals $\mathbb{O}(\mathcal{A}_Q)$, where Q is the quiver geometrically defined in Section 3. That is, we construct an inclusion $\mathbb{O}(\mathfrak{M}(\Lambda, T)) \subseteq \mathbb{O}(\mathcal{A}_Q)$ and show that it is an equality. We also provide symplectic geometric meaning to the $\mathcal{A}$–variables in Section 4. In conjunction with [Casals et al. 2022] — which logically depends on [Shen and Weng 2021] and the present manuscript — we know that $\mathbb{O}(\mathcal{A}_Q) = \mathbb{A}_Q$. Therefore, $\mathfrak{M}(\Lambda, T)$ admits a cluster structure in the strongest possible sense: it is an affine scheme whose ring of regular functions *equals* the upper cluster algebra $\mathbb{O}(\mathcal{A}_Q)$, and also the cluster algebra $\mathbb{A}_Q$.

Remark 2.16 Similarly, a consequence of our results is that $\mathbb{O}(\mathcal{M}_1(\Lambda))$ equals $\mathbb{O}(\mathcal{X}_Q)$, which was also an open question. That is, we show that the inclusion $\mathbb{O}(\mathcal{M}_1(\Lambda)) \subseteq \mathbb{O}(\mathcal{X}_Q)$ is an equality. The results of [Shende et al. 2019], when combined with their later work [Shende et al. 2016], would likely imply that

for Λ associated to plabic fence (no lollipops) one has the inclusion $\mathbb{O}(\mathcal{M}_1(\Lambda)) \subset \mathbb{O}(\mathcal{X}_Q)$. That said, even [Shende et al. 2016; 2019] combined do not prove the equality in these cases.

Finally, we emphasize that the geometric description of the cluster $\mathscr{A}$–variables $A_i \in \mathbb{A}_Q$ and the particular algebraic geometric description of $\mathfrak{M}(\Lambda, T)$ is what allows for the equalities $\mathbb{O}(\mathfrak{M}(\Lambda, T)) = \mathbb{O}(\mathscr{A}_Q)$ to be proven in Section 4. In particular, the fact that $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ is a unique factorization domain (Section 4.2) and the fact that the (candidate) cluster $\mathscr{A}$–variables are irreducible in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ (Section 4.9) are key to deduce $\mathbb{O}(\mathfrak{M}(\Lambda, T)) = \mathbb{O}(\mathscr{A}_Q)$.

3 Diagrammatic weave calculus and initial cycles

The new machinery from contact topology that allows us to construct cluster structures is the study of Legendrian weaves, as initiated in [Casals and Zaslow 2022]. We continue to develop techniques for Legendrian weaves so as to prove Theorem 1.1. These new weave techniques now relate to GP graphs $\mathbb{G}$ and their associated Legendrian links $\Lambda(\mathbb{G})$. Among many central facts, the construction of a weave $\mathfrak{w}(\mathbb{G})$ associated to $\mathbb{G}$ yields a canonical embedded exact Lagrangian filling for $\Lambda(\mathbb{G})$, a sheaf quantization, and the flag moduli of the weaves $\mathfrak{w}(\mathbb{G})$ shall provide the initial seeds for our cluster structures. In addition, as explained in Section 4, the weave $\mathfrak{w}(\mathbb{G})$ is used also to carry the explicit computations necessary for the study of cluster $\mathscr{A}$–variables and the proof of Theorem 1.1.

3.1 Preliminaries on weaves

The reader is referred to [Casals and Zaslow 2022] for the details and background on Legendrian weaves, but we provide here a quick primer on the basics. Let $J, K \subset \mathbb{R}^2$ be two trivalent planar graphs having an isolated intersection point at a common vertex $v \in J \cap K$. By definition, the intersection v is said to be *hexagonal* if the six half-edges in C incident to v interlace, ie alternately belong to J and K. Figure 11, right, depicts such a hexagonal vertex.

Definition 3.1 Given $n \in \mathbb{N}$, an N–weave $\mathfrak{w} \subset \mathbb{R}^2$ is a set $\mathfrak{w} = \{G_i\}_{1 \le i \le N-1}$ of $N-1$ embedded trivalent planar graphs $G_i \subset \mathbb{R}^2$, possibly empty or disconnected, such that G_i is allowed to intersect G_{i+1} only at hexagonal points for $1 \le i \le N - 2$. By definition, a weave $\mathfrak{w} \subset \mathbb{R}^2$ is an N–weave for some $N \in \mathbb{N}$.
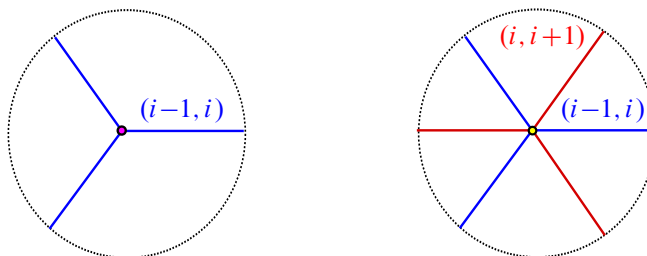
Figure 11: Trivalent vertex (left) and hexagonal point (right).

We also refer to the image of a weave in the plane as a weave $\mathfrak{w}$, as no confusion arises. The edges of the graphs that constitute a weave $\mathfrak{w}$ are often referred to as weave lines. We note that two graphs $G_i, G_j \subset \mathbb{R}^2$ are allowed to intersect (anywhere) as long as $j \neq i, i \pm 1$, and we always assume that the intersection is transverse. Through its image, we also think of an N–weave as an immersed graph in $\mathbb{R}^2$ with colored (or labeled) edges, the color i corresponding to the graph G_i for $1 \leq i \leq N - 1$. Edges labeled by numbers differing by two or more may pass through one another (hence the immersed property, which is met generically), but not at a vertex. As a graph in the plane, an N–weave has trivalent, tetravalent and hexagonal vertices.

Let $\{s_i\}_{i=1}^{N-1}$ be the set of Coxeter generators of the symmetric group S_N. Instead of colors, we can equivalently label the edges of an N–weave $\mathfrak{w} = \{G_i\}$ which belong to the graph G_i with the transposition s_i: these labeled edges will also be referred to as s_i–edges, or i–edges. The theory of weaves as developed in [Casals and Zaslow 2022] is grounded on the theory of Legendrian surfaces in $(\mathbb{R}^5, \xi_{\mathrm{st}})$ and their spatial wavefronts. In brief, a weave $\mathfrak{w} \subset \mathbb{R}^2$ gives rise to a spatial Legendrian wavefront $\Sigma(\mathfrak{w}) \subset \mathbb{R}^3$, which itself lifts to an embedded Legendrian surface $\Lambda(\mathfrak{w})$ in $(\mathbb{R}^5, \xi_{\mathrm{st}})$. The main property of the surface $\Lambda(\mathfrak{w})$ that we use here is that its image $L(\mathfrak{w}) := \pi(\Lambda(\mathfrak{w})) \subset (\mathbb{R}^4, \omega_{\mathrm{st}})$ is an exact Lagrangian surface in the standard symplectic Darboux ball, where $\pi : (\mathbb{R}^5, \xi_{\mathrm{st}}) \to (\mathbb{R}^4, \omega_{\mathrm{st}})$ is the projection along the α_{st}–Reeb direction.

Unless it is stated otherwise, all the weaves that we construct are free — see [Casals and Zaslow 2022, Section 7.1.2] — which translates into the fact that $L(\mathfrak{w}) \subset (\mathbb{R}^4, \omega_{\mathrm{st}})$ will always be an embedded exact Lagrangian surface, and not just immersed. In particular, this implies that $L(\mathfrak{w})$ must have boundary, which it always will. Moreover, when $\mathfrak{w}$ is free, the Lagrangian projection map π is a *homeomorphism*, and hence $H_1(L(\mathfrak{w})) \cong H_1(\Lambda(\mathfrak{w}))$. The underlying contact geometry dictates that certain weaves ought to be considered equivalent. This leads to the following:

Definition 3.2 The moves depicted in Figure 12 are referred to as *weave equivalences*. By definition, two weaves $G, G' \subset \mathbb{R}^2$ are said to be equivalent if they differ by a sequence of weave equivalences or diffeomorphisms of the plane. We interchangeably refer to a weave and its weave equivalence class when the context permits.
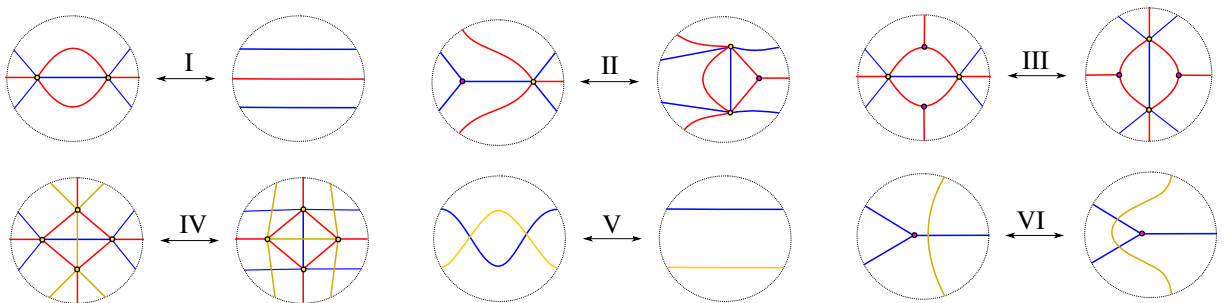
Figure 12: Six weave equivalences.
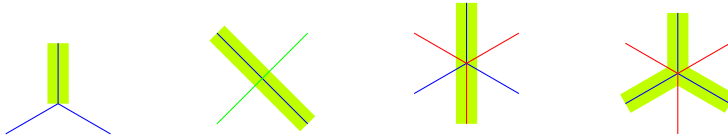
Figure 13: The local models for a Y–tree. The Y–tree is highlighted in light green.

Remark 3.3 As noted in [Casals and Zaslow 2022, Section 4], these moves are not entirely independent, and Move III can be deduced from Move I and Move II. It is nevertheless useful to underscore Move III when working with weaves. The results in [Casals and Zaslow 2022], using the underlying contact geometry, imply that all the constructions that we associate to a weave are invariant under weave equivalences. (It would to be possible to verify this combinatorially as well; see for instance the computations in [Casals et al. 2020].)

A first goal is constructing a weave for each GP graph $\mathbb{G} \subset \mathbb{R}^2$; it is achieved in Section 3.3 once we have reviewed the necessary material on Y–cycles and weave mutation.

3.2 Y–cycles and weave mutation

Let $\mathfrak{w} \subset \mathbb{R}^2$ be a free weave. The homology group $H_1(L(\mathfrak{w})) \cong H_1(\Lambda(\mathfrak{w}))$ has a central role in our article, as it is a sublattice of the defining lattice for the initial seed. Casals and Zaslow [2022, Section 2] devised a method to describe absolute cycles in $L(\mathfrak{w})$ in terms of $\mathfrak{w} \subset \mathbb{R}^2$. The main concept that is relevant for our purposes is that of a Y–cycle on a weave $\mathfrak{w}$, which is defined as follows:

Definition 3.4 Let $\mathfrak{w} \subset \mathbb{R}^2$ be a weave. An absolute 1–cycle $\gamma \subset \Lambda(\mathfrak{w})$ is said to be a Y–*cycle* if its projection onto $\mathbb{R}^2$ consists of weave lines, ie it is contained in $\mathfrak{w}$. A Y–cycle is said to be a Y–*tree* if its projection image is a tree, considered as a planar embedded graph in $\mathbb{R}^2$. A Y–tree is a l–*cycle* if its projection onto $\mathbb{R}^2$ does not have any trivalent vertices. Finally, an l–cycle is *short* if it does not pass through any hexagonal vertex of the weave $\mathfrak{w}$. Figure 13 depicts the four possible local models for a Y–tree near a trivalent, tetravalent, and hexagonal vertex of the weave.

Definition 3.4 allows us to associate a unique absolute cycle on $\Lambda(\mathfrak{w})$ (and hence on $L(\mathfrak{w})$) to each Y–tree in a weave $\mathfrak{w}$, as explained in [Casals and Zaslow 2022, Section 2]. (There are two conventions regarding orientations and choice of sheet at which to lift, but, once these conventions are fixed, the absolute cycle is defined uniquely.) Note that a Y–cycle can stack multiple copies of the above patterns at the same vertex; when this happens at a trivalent or hexagonal vertex, the stacking creates self-intersections of the absolute cycle it represents. The distinction between *embedded* and *immersed* representatives of absolute homology classes is at the core of the distinction between *mutable* and *frozen* variables for the cluster structures we construct. The outstanding role of Y–trees is justified by the following fact:
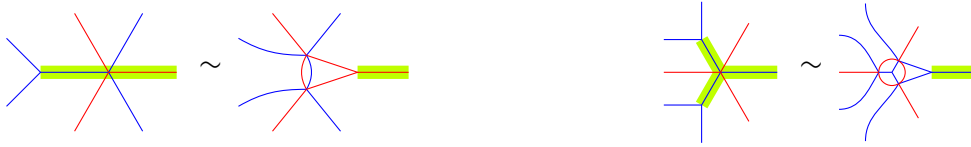
Figure 14: Local weave equivalences to turn a Y–tree into a short I–cycle. Note that the first row is just Move II from Figure 12, where we kept track of the Y–tree — highlighted in light green — before and after the equivalence.

Proposition 3.5 *Let $\mathfrak{w} \subset \mathbb{R}^2$ be a free weave and δ be an absolute 1–cycle representing a homology class in $H_1(L(\mathfrak{w})) \cong H_1(\Lambda(\mathfrak{w}))$ which is obtained from a Y–tree in $\mathfrak{w}$. Then there exists a weave equivalence $\mathfrak{w} \sim \mathfrak{w}'$ such that the cycle $\delta \subset \mathfrak{w}$ becomes a short I–cycle in $\mathfrak{w}'$. In consequence, any homology class in $H_1(L(\mathfrak{w}))$ represented by a Y–tree admits an embedded representative $\gamma \subset L(\mathfrak{w})$ which bounds an embedded exact Lagrangian disk in $\mathbb{R}^4 \setminus L(\mathfrak{w})$.*

Proof This readily follows from [Casals and Zaslow 2022], by applying the equivalence moves in Figure 12 and keeping track of the change of a Y–tree under these moves. In fact, it suffices to use of the two local weave equivalences shown in Figure 14.

By using the two weave equivalences in Figure 14, we can work outside in on the Y–tree δ and replace each weave line of δ with a double track, and shorten δ to a short I–cycle somewhere along the original Y–tree. The double tracks that appear in this shortening process are schematically depicted in Figure 15. The second half of the proposition follows from the description of a short I–cycle in [Casals and Zaslow 2022].   □

Remark 3.6 Proposition 3.5 implies that any one Y–tree can be turned into a short I–cycle after weave equivalences. It is not the case that a Y–cycle, which is not necessarily a Y–tree, can always be turned into a short I–cycle. It is also not the case that Proposition 3.5 works for more than one Y–tree at once, in the following sense. If two Y–trees $Y_1, Y_2 \subset \mathfrak{w}$ are given in a weave $\mathfrak{w}$, then there exists a sequence of weave equivalences from $\mathfrak{w}$ to a weave $\mathfrak{w}_1$ such that Y_1 becomes a short I–cycle in $\mathfrak{w}_1$. There is no guarantee that Y_2 will be a short I–cycle in $\mathfrak{w}_1$; Y_2 will be a short I–cycle in another weave $\mathfrak{w}_2$ equivalent to $\mathfrak{w}$, a priori different from $\mathfrak{w}_1$. More generally, there are collections of Y–trees in a weave $\mathfrak{w}$ such that
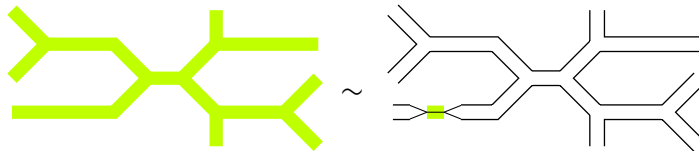
Figure 15: Left: a Y–tree cycle highlighted in light green. Right: the double tracks that remain on the (equivalent) weave after the shortening process, where the Y–cycle has now become the short I–cycle drawn in light green.
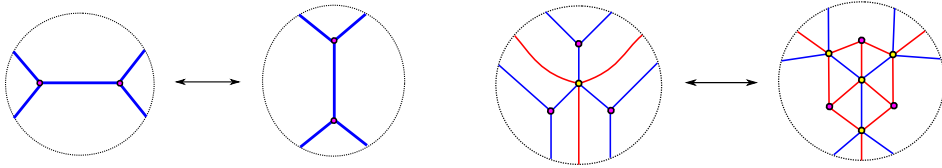
Figure 16: Left: weave mutation along the short l–cycle given by the blue edge. Right: a weave mutation along a monochromatic Y–tree.

there is no sequence of weave equivalences that would turn at once all those Y–trees in $\mathfrak{w}$ into short l–cycles in any one weave $\mathfrak{w}'$ equivalent to $\mathfrak{w}$.

The existence of the embedded Lagrangian disks from Proposition 3.5, ie $\mathbb{L}$–compressible curves in $L(\mathfrak{w})$, allows us to perform Lagrangian disk surgeries along Y–trees and produce new exact Lagrangian fillings. Casals and Zaslow [2022, Section 4.8] proved that it is possible to describe this symplectic geometric operation via a diagrammatic change in a piece of the weave called "weave mutations". This leads to the following definition:

Definition 3.7 Let $\gamma \subset \mathfrak{w}$ be the short l–cycle — a monochromatic blue edge — depicted in Figure 16, left. Then the local move illustrated in Figure 16, left, is said to be a *weave mutation* at the short l–cycle γ. This is the standard Whitehead move for trivalent graphs, dual to a flip in a triangulation. Note that a weave mutation replaces the short l–cycle γ with a new short l–cycle, which we often denote by γ'; we call γ' the *image* of γ under the weave mutation.

Definition 3.8 For a Y–tree in general, one can apply Proposition 3.5 to turn it into a short l–cycle, perform a weave mutation, and then apply some other weave equivalences. Thus, a *weave mutation* at a Y–tree γ in general means a weave mutation at its short l–cycle counterpart conjugated by sequences of weave equivalences. By following the sequences of weave equivalences, the weave mutation replaces γ by its *image*, which is a new Y–tree γ'.

Definition 3.9 Two weaves $\mathfrak{w}$ and $\mathfrak{w}'$ are said to be (weave) *mutation equivalent* if they can be connected by a sequence of moves consisting of weave equivalences and weave mutations.

Finally, we emphasize that weave mutations will allow us to mutate at Y–trees of $\mathfrak{w}(\mathbb{G})$ corresponding to faces and regions of $\mathbb{G}$, even if they might not be square. The resulting weave, typically, will not be of the form $\mathfrak{w}(\mathbb{G}')$ for any GP graph $\mathbb{G}'$, but all the diagrammatic and symplectic geometric results developed in this article and [Casals and Zaslow 2022] can still be applied.

3.3 Initial weave for a GP graph

In this section we construct the initial weave $\mathfrak{w}(\mathbb{G})$ associated to a GP graph $\mathbb{G}$. This is done by breaking $\mathbb{G}$ into elementary columns and assigning a local weave associated to each such column. Recall

the standard reduced word $\mathtt{w}_{0,n} = s_{[1,1]}s_{[2,1]}s_{[3,1]} \ldots s_{[n-1,1]}$ for the longest element $w_{0,n} \in S_n$. Let us define $\ell = \ell(\mathtt{w}_{0,n}) = \frac{1}{2}n(n-1)$. The first three local weaves $\mathfrak{n}(w), \mathfrak{c}^\uparrow(w)$ and $\mathfrak{c}^\downarrow(w)$ are defined as follows:

Definition 3.10 Let $w = s_{i_1} \ldots s_{i_\ell}$ be a reduced expression for $w_{0,n} \in S_n$. By definition, the weave $\mathfrak{n}(w)$ is given by n horizontal parallel weave lines such that the jth strand, counting from the bottom, is labeled by the transposition s_{i_j} for $j \in [1, \ell]$.

The weave $\mathfrak{c}^\uparrow(w)$ is given by the weave $\mathfrak{n}(w)$ where a trivalent vertex is added at the top strand — labeled by s_{i_ℓ} — such that the third leg of this trivalent vertex is a vertical ray starting at the top strand and continuing upwards.

Similarly, the weave $\mathfrak{c}^\downarrow(w)$ is given by the weave $\mathfrak{n}(w)$ where a trivalent vertex is added at the bottom strand — labeled by s_{i_1} — such that the third leg of this trivalent vertex is a vertical ray starting at the bottom strand and continuing downwards.

As explained above, the weave $\mathfrak{w}(\mathbb{G})$ associated to $\mathbb{G}$ is built by horizontally concatenating weaves local models: each local model is associated to one of the three types of elementary columns. The corresponding weaves for each of these occurrences are described as follows:

3.3.1 Local weaves for Type 1 columns

Definition 3.11 (weave for Type 1) The weave associated to a Type 1 column of a GP graph $\mathbb{G}$, which consists of n parallel horizontal lines, is $\mathfrak{n}(\mathtt{w}_{0,n})$, where $\mathtt{w}_{0,n}$ is the standard reduced expression for the longest element in a symmetric group S_k.

It is important to note that the transpositions s_i labeling the strands of $\mathfrak{n}(\mathtt{w}_{0,n})$ depend on the simple transpositions that generate the symmetric group $S_{[a,b]}$ associated to that specific region (see Section 2.1). Due to the appearance of lollipops in the GP graph, the different symmetric groups $S_{[a,b]}$ that we encounter (as we read the GP graph left to right) have varying discrete intervals $[a, b]$.

3.3.2 Local weaves for Type 2 columns It is a well-known property of the symmetric group — see for instance [Björner and Brenti 2005, Section 3.3] — that any two reduced word expressions for the same element can be transformed into each other via finite sequences of the following two moves:

- $s_i s_j \sim s_j s_i$ if $|i - j| > 1$.
- $s_i s_j s_i \sim s_j s_i s_j$ if $|i - j| = 1$.

Now consider the weave $\mathfrak{n}(\mathtt{w}_{0,n})$ and the s_1–strand labeled by the ith appearance of s_1 in the standard reduced expression $\mathtt{w}_{0,n}$. In order to construct the weave for a Type 2 column, in Definition 3.15 below, we need the following auxiliary local weaves:
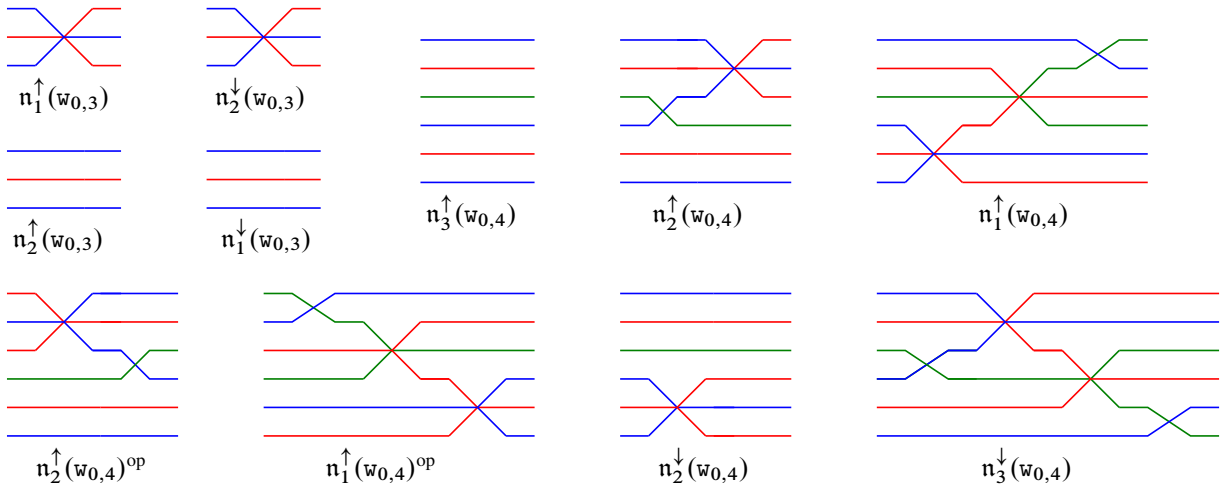
Figure 17: Instances of the weaves $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})$ and $\mathfrak{n}_i^\downarrow(\mathtt{w}_{0,n})$ and their opposites, from Definition 3.12, in some of the cases for $n = 3, 4$.

Definition 3.12 The weave $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})$ is the unique horizontal weave that coincides with $\mathfrak{n}(\mathtt{w}_{0,n})$ at the left, contains only tetravalent and hexagonal vertices, and brings the ith s_1–strand of $\mathtt{w}_{0,n}$ to the top level, using a minimal number of weave vertices.

Similarly, the weave $\mathfrak{n}_i^\downarrow(\mathtt{w}_{0,n})$ is the unique weave that coincides with $\mathfrak{n}(\mathtt{w}_{0,n})$ at the left, contains only tetravalent and hexagonal vertices, and brings the ith s_1–strand of $\mathtt{w}_{0,n}$ to the bottom level, using a minimal number of weave vertices.

Finally, we denote by $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})^{\text{op}}$ and $\mathfrak{n}_i^\downarrow(\mathtt{w}_{0,n})^{\text{op}}$ the weaves obtained by reflecting $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})$ and $\mathfrak{n}_i^\downarrow(\mathtt{w}_{0,n})$ along a (disjoint) vertical axis.

In Definition 3.12, bringing the ith s_1–strand of $\mathtt{w}_{0,n}$ to the top level means considering a horizontal weave which starts at $\mathfrak{n}(\mathtt{w}_{0,n})$ on the left-hand side and contains a sequence of tetravalent and hexagonal vertices (no trivalent vertices) such that following the ith s_1–strand of $\mathtt{w}_{0,n}$ under these vertices (passing through them straight) ends up at the top strand at the right-hand side. There are many weaves that verify this property, but, by the Zamolodchikov relation proven in [Casals and Zaslow 2022], they are all equivalent and we might as well take the one with a minimal number of vertices. Figure 17 illustrates several examples of the weaves $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})$ and $\mathfrak{n}_i^\downarrow(\mathtt{w}_{0,n})$ in Definition 3.12 for $n = 3, 4$. Note that $\mathfrak{n}_{n-1}^\uparrow(\mathtt{w}_{0,n}) = \mathfrak{n}_{n-1}^\uparrow(\mathtt{w}_{0,n})^{\text{op}} = \mathfrak{n}_1^\downarrow(\mathtt{w}_{0,n}) = \mathfrak{n}_1^\downarrow(\mathtt{w}_{0,n})^{\text{op}} = \mathfrak{n}(\mathtt{w}_{0,n})$ for any $n \in \mathbb{N}$.

Definition 3.13 The weave $\mathfrak{c}_i^\uparrow(\mathtt{w}_{0,n})$ is the weave obtained by horizontally concatenating the three weaves $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})$, $\mathfrak{c}^\uparrow(w_i)$, and $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})^{\text{op}}$, left to right, where w_i denotes the reduced expression for $\mathtt{w}_{0,n}$ found at the right of the weave $\mathfrak{n}_i^\uparrow(\mathtt{w}_{0,n})$.
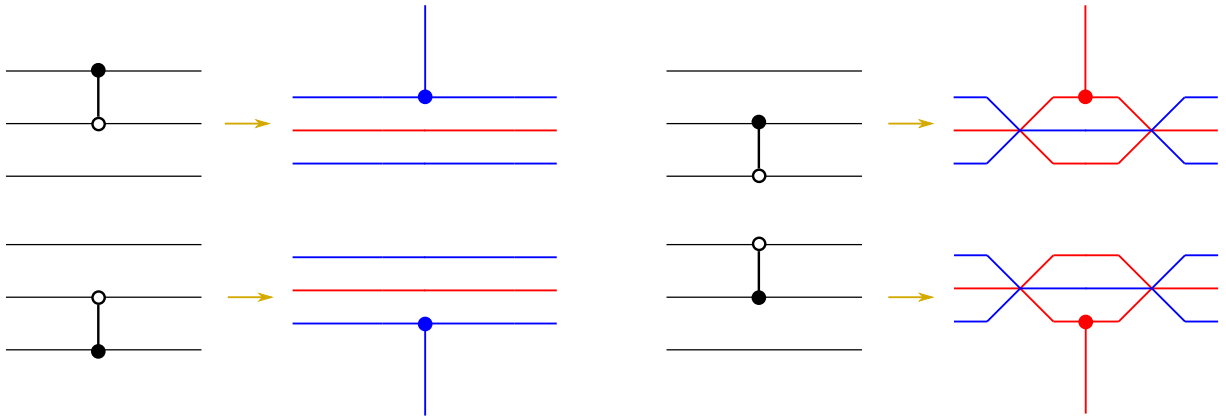
Figure 18: All possible Type 2 columns for $n = 3$ and their corresponding weaves. In detail, $\mathfrak{c}_2^{\uparrow}(\mathsf{w}_{0,3})$ (upper left), $\mathfrak{c}_1^{\uparrow}(\mathsf{w}_{0,3})$ (upper right), $\mathfrak{c}_1^{\downarrow}(\mathsf{w}_{0,3})$ (lower left) and $\mathfrak{c}_2^{\downarrow}(\mathsf{w}_{0,3})$ (lower right).

Definition 3.14 Similarly, the weave $\mathfrak{c}_i^{\downarrow}(\mathsf{w}_{0,n})$ is the weave obtained by horizontally concatenating the three weaves $\mathfrak{n}_i^{\downarrow}(\mathsf{w}_{0,n})$, $\mathfrak{c}^{\downarrow}(w_i)$ and $\mathfrak{n}_i^{\downarrow}(\mathsf{w}_{0,n})^{\mathrm{op}}$, left to right, where w_i denotes the reduced expression for $\mathsf{w}_{0,n}$ found at the right of the weave $\mathfrak{n}_i^{\downarrow}(\mathsf{w}_{0,n})$.

Figure 18 illustrates examples of the weaves $\mathfrak{c}_i^{\uparrow}(\mathsf{w}_{0,n})$ and $\mathfrak{c}_i^{\downarrow}(\mathsf{w}_{0,n})$ in Definitions 3.13 and 3.14 for $n = 3$. For the next definition, recall that we always label the horizontal lines, in a Type 1 or Type 2 column of the GP graph $\mathbb{G}$, with consecutive natural numbers, from bottom to top. For the moment, let us assume that these labels are in $[1, n]$.

Definition 3.15 (weave for Type 2) For $i \in [1, n-1]$, the weave associated to a Type 2 column of a GP graph $\mathbb{G}$ whose vertical edge has a white vertex at the ith horizontal line and a black vertex at the $(i+1)^{\text{st}}$ horizontal line is the weave $\mathfrak{c}_i^{\uparrow}(\mathsf{w}_{0,n})$, and the weave associated to a Type 2 column of a GP graph $\mathbb{G}$ whose vertical edge has a black vertex at the ith horizontal line and a white vertex at the $(i+1)^{\text{st}}$ horizontal line is the weave $\mathfrak{c}_{n-i}^{\downarrow}(\mathsf{w}_{0,n})$.

3.3.3 Local weaves for Type 3 columns Let us consider a column of Type 3 with labels $1, 2, \ldots, n$ for the horizontal lines on the right (counting from bottom to top) and with a white lollipop attached to the ith horizontal line with $i \in [1, n]$. The case of a black lollipop is similar, and discussed later.

By construction, the weaves associated with the two Type 1 columns sandwiching this Type 3 column are $\mathfrak{n}(\mathsf{w}_{0,n-1})$ and $\mathfrak{n}(\mathsf{w}_{0,n})$, respectively. Hence, the weave we associate to such a Type 3 column must have these boundary conditions. Let us start with the following weave:

Definition 3.16 The weave $\mathfrak{i}_i^w$ is the unique weave with no weave vertices, satisfying:

(i) At its left, $\mathfrak{i}_i^w$ coincides with the horizontal weave $\mathfrak{n}(\mathsf{w}_{0,n-1})$, and at its right, $\mathfrak{i}_i^w$ coincides with the horizontal weave $\mathfrak{n}(s_{[n-1,i]}^{-1} \mathsf{w}_{0,n-1} s_{[n-1,n-i+1]})$.
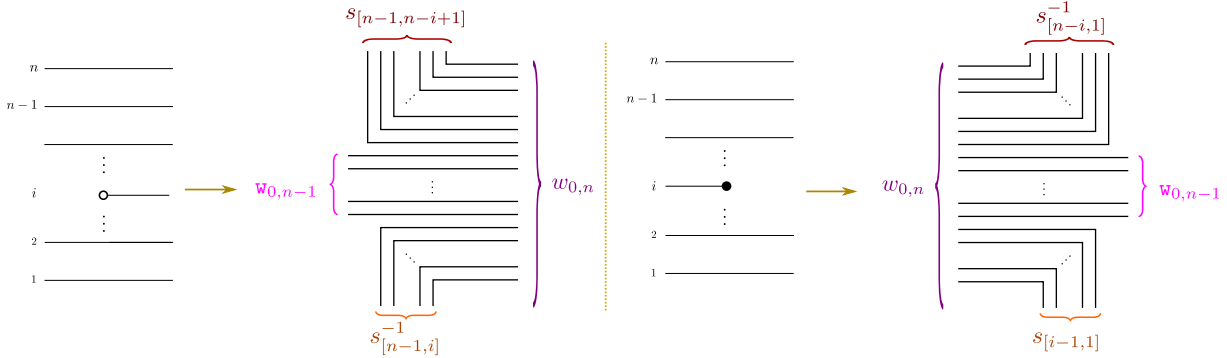
Figure 19: The weaves $\mathfrak{i}_i^w$ (left) and $\mathfrak{i}_i^b$ (right) from Definitions 3.16 and 3.18.

(ii) The weave lines in $\mathfrak{n}(s_{[n-1,i]}^{-1} w_{0,n-1} s_{[n-1,n-i+1]})$ labeled by the transpositions in the reduced expression $s_{[n-1,n-i+1]}$ diverge upwards to vertical rays.

(iii) The weave lines in $\mathfrak{n}(s_{[n-1,i]}^{-1} w_{0,n-1} s_{[n-1,n-i+1]})$ labeled by the transpositions in the reduced expression $s_{[n-1,i]}^{-1}$ diverge downwards to vertical rays.

See Figure 19, left, for a depiction of $\mathfrak{i}_i^w$, illustrating what is meant by diverging upwards and downwards to vertical rays.

Note that the word $\mathfrak{n}(s_{[n-1,i]}^{-1} w_{0,n-1} s_{[n-1,n-i+1]})$ in Definition 3.16 is a reduced expression for the half-twist $w_{0,n}$. Now, the weaves $\mathfrak{i}_i^w$ in Definition 3.16 cannot quite be the weaves for the Type 3 column yet because the labeling on the right-hand side is not $w_{0,n}$, but rather $\mathfrak{n}(s_{[n-1,i]}^{-1} w_{0,n-1} s_{[n-1,n-i+1]})$. To fix this, let $\mathfrak{n}_i^w$ be any horizontal weave that coincides with $\mathfrak{n}(s_{[n-1,i]}^{-1} w_{0,n-1} s_{[n-1,n-i+1]})$ on the left, coincides with $\mathfrak{n}(w_{0,n})$ on the right, and with no trivalent weave vertices in the middle. Any choice of $\mathfrak{n}(w_{0,n})$ would yield an equivalent weave.

Definition 3.17 (weave for white lollipop) The weave $\mathfrak{l}_i^w$ associated to a Type 3 column with a white lollipop at the ith horizontal line is the horizontal concatenation of $\mathfrak{i}_i^w$ and $\mathfrak{n}_i^w$.

Figures 20 and 21 illustrate the weaves $\mathfrak{l}_1^w, \mathfrak{l}_2^w, \mathfrak{l}_3^w$ and $\mathfrak{l}_4^w$ for $n = 4$, with the coloring convention that s_1 is blue, s_2 is red and s_3 is green. The pink boxes in the figures contain the $\mathfrak{i}_i^w$ pieces, and the yellow boxes contain the $\mathfrak{n}_i^w$ pieces. The figures also draw the corresponding pieces of the fronts $\mathfrak{f}(\mathbb{G})$, explaining the contact-geometric origin of these weaves.

The case of a column of Type 3 with labels $1, 2, \ldots, n$ for the horizontal lines on the right, and a black lollipop at the ith horizontal line is similar. The necessary definitions are as follows:

Definition 3.18 The weave $\mathfrak{i}_i^b$ is the unique weave with no weave vertices, satisfying:

(i) At its left, $\mathfrak{i}_i^b$ coincides with the horizontal weave $\mathfrak{n}(s_{[i-1,1]} w_{0,n-1} s_{[n-i,1]}^{-1})$ and, at its right, $\mathfrak{i}_i^b$ coincides with the horizontal weave $\mathfrak{n}(w_{0,n-1})$.
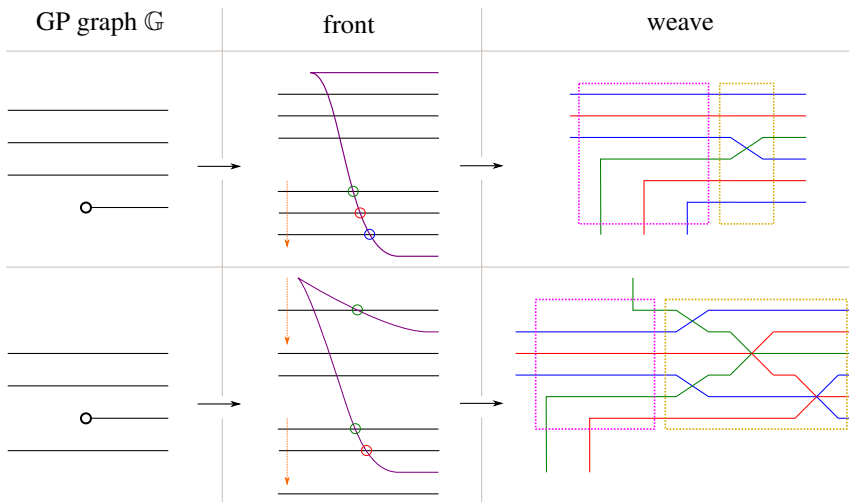
Figure 20: Weaves $\mathfrak{l}_i^w$ associated to a white lollipop with $n = 4$, as in Definition 3.17. The first row depicts the case $i = 1$ and the second row the case $i = 2$. The weaves $\mathfrak{n}_i^w$ are drawn within the yellow boxes. The weaves $\mathfrak{i}_i^w$, with the incoming weave strands, are depicted within the pink boxes.

(ii) The weave lines in $\mathfrak{n}(s_{[i-1,1]}\mathtt{w}_{0,n-1}s_{[n-i,1]}^{-1})$ labeled by the transpositions in the reduced expression $s_{[i-1,1]}$ diverge downwards to vertical rays.

(iii) The weave lines in $\mathfrak{n}(s_{[i-1,1]}\mathtt{w}_{0,n-1}s_{[n-i,1]}^{-1})$ labeled by the transpositions in the reduced expression $s_{[n-i,1]}^{-1}$ diverge upwards to vertical rays.
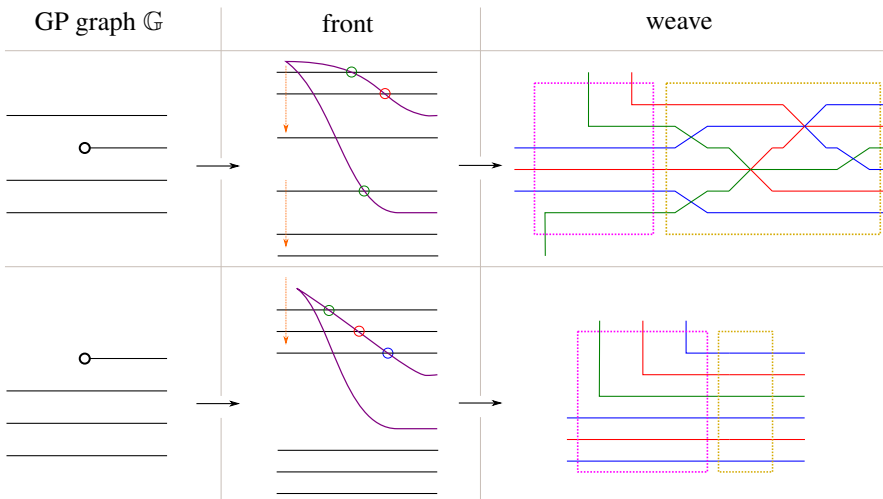
Figure 21: Weaves $\mathfrak{l}_i^w$ associated to a white lollipop with $n = 4$. The first row depicts the case $i = 3$ and the second row the case $i = 4$. The weaves $\mathfrak{n}_i^w$ are drawn within the yellow boxes, and the weaves $\mathfrak{i}_i^w$ are depicted in the pink boxes.
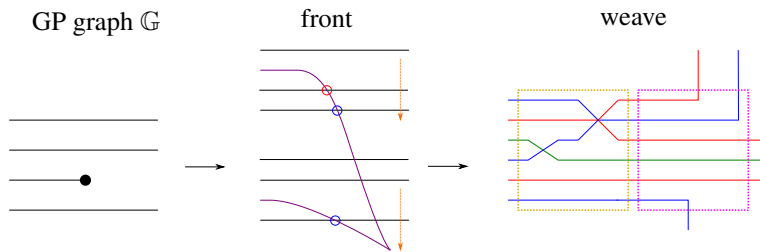
Figure 22: The weave $\mathfrak{l}_2^b$ associated to a black lollipop when $n = 4$, in accordance with Definition 3.19. The weave $\mathfrak{n}_2^b$ is drawn in the yellow box, and $\mathfrak{i}_2^b$ in the pink box. Note that the departing strands of the weave (in the pink box) are in bijection with the crossings of the front, as indicated.

See Figure 19, right, for a depiction of $\mathfrak{i}_i^b$. Similarly, we denote by $\mathfrak{n}_i^b$ any horizontal weave which coincides with the horizontal weave $\mathfrak{n}(s_{[i-1,1]}\mathsf{w}_{0,n-1}s_{[n-i,1]}^{-1})$ on the right, coincides with $\mathfrak{n}(\mathsf{w}_{0,n})$ on the left, and with no trivalent weave vertices in the middle.

Definition 3.19 (weave for black lollipop) The weave $\mathfrak{l}_i^b$ associated to a Type 3 column with a black lollipop at the ith horizontal line is the horizontal concatenation of $\mathfrak{n}_i^b$ and $\mathfrak{i}_i^b$.

See Figure 22 for an example of $\mathfrak{l}_i^b$ in the case $i = 2$ and $n = 4$.

3.3.4 The initial weave Let $\mathbb{G}$ be a GP graph. Sections 3.3.1, 3.3.2 and 3.3.3 have explained how to obtain a weave from each of the three types of elementary columns. Note that each of these weaves coincides with $\mathfrak{n}(\mathsf{w}_{0,k})$ and with $\mathfrak{n}(\mathsf{w}_{0,m})$ for some k and m at its two ends. For Type 1 and Type 2, $k = m$, and, for Type 3, $|k - m| = 1$. Note that, if we consider two adjacent elementary columns in $\mathbb{G}$, the associated weaves will coincide at the common side, and thus can be horizontally concatenated.

Definition 3.20 (initial weave) Let $\mathbb{G}$ be a GP graph. The initial weave $\mathfrak{w}(\mathbb{G})$ associated to $\mathbb{G}$ is the weave obtained by subdividing $\mathbb{G}$ into elementary columns and then concatenating the weaves associated with each elementary column, in the order dictated by the columns.

3.4 Topology of the initial weave

Let $\mathbb{G}$ be a GP graph and $\mathfrak{w}(\mathbb{G}) \subset \mathbb{R}^2$ its initial weave. In this subsection we show how to obtain a Legendrian link $\Lambda(\mathbb{G}) \subset (\mathbb{R}^3, \xi_{\text{st}})$ and an embedded exact Lagrangian filling $L(\mathfrak{w}) \subset (\mathbb{R}^4, \lambda_{\text{st}})$ from the weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$.

3.4.1 The braid of an initial weave Suppose $\mathfrak{w}(\mathbb{G})$ is an N–weave. Let $K \subset \mathbb{R}^2$ be a compact subset such that $(\mathbb{R}^2 \setminus K) \cap \mathfrak{w}(\mathbb{G})$ contains no weave vertices. Note that the number of weave lines in $(\mathbb{R}^2 \setminus K) \cap \mathfrak{w}(\mathbb{G})$ and their labeling is independent of any K with this property. The weave lines are labeled by simple transpositions $s_i \in S_N$, which can be lifted to unique positive generators σ_i in the Artin braid group Br_N.

Definition 3.21 Let $\beta(\mathbb{G})$ be the positive braid word obtained by reading the positive braid generators associated with the weave lines of $(\mathbb{R}^2 \setminus K) \cap \mathfrak{w}(\mathbb{G})$ in a counterclockwise manner, starting at the unique strand the corresponds to the leftmost white lollipop in $\mathbb{G}$.

Part of the usefulness of Definition 3.21 is the following simple lemma:

Lemma 3.22 *Let $\mathbb{G}$ be a GP graph, $\mathfrak{w}(\mathbb{G}) \subset \mathbb{R}^2$ its initial weave and $\beta(\mathbb{G})$ its positive braid word. Then the (-1)–framed closure of $\beta(\mathfrak{w}(\mathbb{G}))$ is a front for the Legendrian link $\Lambda(\mathbb{G})$.*

Lemma 3.22 can be phrased as follows. Consider the Legendrian link $\Lambda(\beta(\mathbb{G})) \subset (\mathbb{R}^3, \xi_{\mathrm{st}})$ whose front is the (-1)–framed closure of the braid word $\beta(\mathbb{G})$. Then the Legendrian links $\Lambda(\beta(\mathbb{G}))$ and $\Lambda(\mathbb{G})$ are Legendrian isotopic in $(\mathbb{R}^3, \xi_{\mathrm{st}})$.

3.4.2 The surface of the initial weave Let $\mathfrak{w} \subset \mathbb{R}^2$ be a weave and $\Lambda(\mathfrak{w}) \subset (\mathbb{R}^5, \xi_{\mathrm{st}})$ the Legendrian represented by its front. By definition, the Lagrangian $L(\mathfrak{w}) \subset (\mathbb{R}^4, \lambda_{\mathrm{st}})$ is the Lagrangian projection of $\Lambda(\mathfrak{w})$. We refer to [Casals and Zaslow 2022, Section 7.1] for details on how weaves yield exact Lagrangian fillings of Legendrian links in $(\mathbb{R}^3, \xi_{\mathrm{st}})$, and recall that $\mathfrak{w}$ is said to be free if $L(\mathfrak{w}) \subset (\mathbb{R}^4, \lambda_{\mathrm{st}})$ is embedded. The following lemma is readily proven:

Lemma 3.23 *Let $\mathbb{G}$ be a GP graph and suppose its initial weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ is an N–weave. Then $L = L(\mathfrak{w}) \subset (\mathbb{R}^4, \lambda_{\mathrm{st}})$ is an embedded exact Lagrangian filling of $\Lambda(\mathbb{G})$ with Euler characteristic*

$$\chi(L) = N - \#(\textit{trivalent vertices of } \mathfrak{w}) = \#(\textit{horizontal lines in } \mathbb{G}) - \#(\textit{vertical edges in } \mathbb{G}).$$

The number of boundary components of $L(\mathfrak{w}(\mathbb{G}))$ is readily computed from $\beta(\mathbb{G})$: it is given by the number of cycles in the cycle decomposition of the Coxeter projection of $\beta(\mathbb{G})$. Finally, a central feature of weaves is the following: it is possible to draw many weaves which coincide with $\mathfrak{w}(\mathbb{G})$, outside a large enough compact set $K \subset \mathbb{R}^2$, and which represent embedded exact Lagrangian fillings of $\Lambda(\mathbb{G})$. In fact, as explained in Section 3.2, there are some local modifications that we can perform to the weave — *weave mutations* — such that the smooth embedded class of the associated (Lagrangian) surface in $(\mathbb{R}^4, \lambda_{\mathrm{st}})$ remains the same but the Hamiltonian isotopy class typically changes. Square face mutation of a GP graph $\mathbb{G}$ is recovered by weave mutations but, importantly, weave mutations allow for more general mutations, including mutations at nonsquare faces of $\mathbb{G}$ and sugar-free regions. The result of such *weave mutations* applied to $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ is again another weave $\mu(\mathfrak{w})$; it may no longer be of the form $\mathfrak{w}(\mathbb{G}')$ for any GP graph $\mathbb{G}'$, but it is a weave and thus, using the calculus in [Casals and Zaslow 2022], we can manipulate it efficiently and use it to prove the results here. In this process, we need explicit geometric cycles representing generators of the absolute homology $H_1(L(\mathfrak{w}))$. In fact, such geometric cycles lead to the quiver for the initial seed. Thus, we now gear towards understanding how to construct geometric representatives of homology classes using weaves.

3.5 Naive absolute cycles in $L(\mathfrak{w}(\mathbb{G}))$

In this subsection we explain how to find a set of geometric (absolute) cycles on $L = L(\mathfrak{w}(\mathbb{G}))$ which generate $H_1(L)$.

Since the genus of an embedded exact Lagrangian filling is determined by the (maximal) Thurston–Bennequin invariant of its Legendrian boundary, all embedded exact Lagrangian fillings of a given Legendrian link are topologically equivalent as abstract surfaces, ie they have the same genus. In the case of a Legendrian link $\Lambda(\mathbb{G})$ associated with a GP graph, it is readily seen that this is the same abstract topological type as that of the Goncharov–Kenyon conjugate surface $S = S(\mathbb{G})$ [2013]. Since the conjugate surface S deformation retracts back to the GP graph $\mathbb{G}$, it follows that the boundaries of the faces of $\mathbb{G}$ form a basis for the absolute homology groups $H_1(\mathbb{G}) \cong H_1(S) \cong H_1(L)$. This basis, indexed by the faces of $\mathbb{G}$, will be referred to as the *naive basis* of $H_1(L)$.

Remark 3.24 This set of generating absolute cycles is *not* good enough in order to construct cluster structures, nor does its intersection quiver give the correct initial quiver. Thus, these cycles will be referred to as the set of *naive* absolute cycles, and we will perform the necessary corrections in Section 3.7.

In order to proceed geometrically, we would like identify the naive basis elements of $H_1(L)$ as lifts of a specific collection of absolute cycles on the weave front $\Sigma = \Sigma(\mathfrak{w}(\mathbb{G}))$, ideally a collection of Y–cycles on $\mathfrak{w}(\mathbb{G})$ (Definition 3.4). Since any GP graph $\mathbb{G}$ can be decomposed into elementary columns, we can try to build these absolute cycles by concatenating appropriate relative cycles associated with each elementary column.

3.5.1 Local representatives of naive absolute cycles in a Type 1 column

In an elementary Type 1 column of $\mathbb{G}$ with n horizontal lines, there are $n - 1$ faces, ie gaps, between these n horizontal lines. For each of these $n - 1$ gaps, we identify a unique weave line as follows. First, we observe that a cross-section of the weave front Σ associated to a Type 1 column is, by construction, the reduced expression $\mathfrak{w}_{0,n}$ of $w_{0,n}$. In this reduced expression, the lowest Coxeter generator (s_i with the smallest i) appears exactly $n - 1$ times. Second, there is a geometric bijection between these $n - 1$ faces and the $n - 1$ appearances of the lowest Coxeter generator in the reduced expression $\mathfrak{w}_{0,k}$. Indeed, for a face f at a Type 1 column, the intersection of ∂f with a Type 1 column has two connected components, which go along two neighboring horizontal lines, say the jth and the $(j+1)^{\text{st}}$. Since each horizontal line is the deformation retract of a sheet in the weave front Σ, a natural choice of the local weave line representative will be the intersection
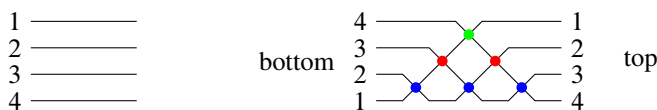
Figure 23: Left: an elementary column with four horizontal lines. Right: the corresponding cross-section for its associated weave surface $\Sigma(\mathbb{G})$.
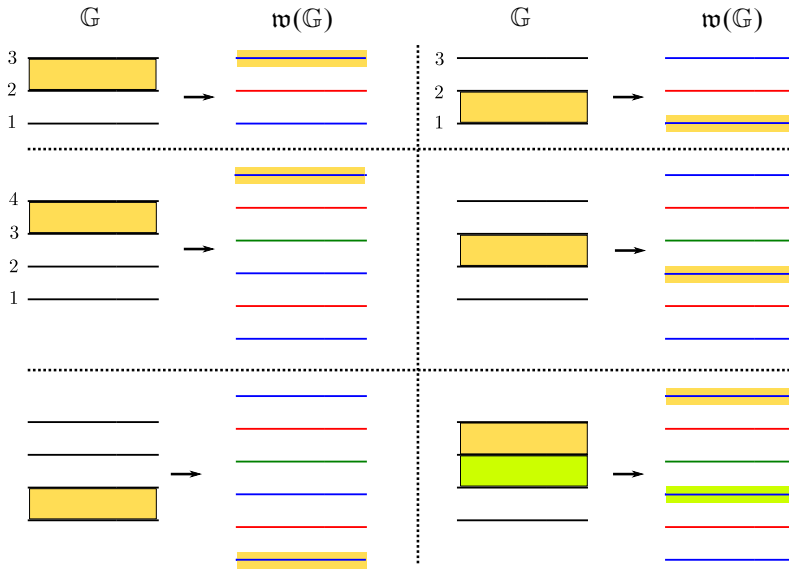
Figure 24: Associating a (piece of a) cycle in the weave for Type 1 columns. The first row depicts the two cases for $n = 3$ strands, and the second and third rows depict the three cases for $n = 4$, and an example with a union, in the lower-right corner. In all cases, the face $f \subset \mathbb{G}$ is highlighted in yellow, and the associated s_1–edge in the weave $\mathfrak{w}(\mathbb{G})$ is also highlighted in the same color. In the last case of the union, in the lower right, one of the faces and its cycle are highlighted in green.

of the two corresponding sheets in Σ, which in turn corresponds to the jth appearance of the lowest Coxeter generator. Figure 23 illustrates a cross-section of the weave front for $n = 4$. Figure 24 illustrates all the possible cases for $n = 3, 4$.

3.5.2 Local representatives of naive absolute cycles in a Type 2 column

Consider a Type 2 column with n horizontal lines and a single vertical edge between the jth and $(j+1)^{\text{st}}$ horizontal lines. First, for the faces bounded by any other pair of consecutive horizontal lines, say the kth and $(k+1)^{\text{st}}$ with $k \neq j$, the associated naive absolute cycle in the weave is the *unique* long I–cycle connecting the corresponding weave cycles on the two adjacent Type 1 columns. In other words, one starts at the kth appearance (counting from below) of the lowest Coxeter generator on the left (see Section 3.5.1) and follows that weave line straight through any hexagonal vertices. By the construction of the weave $\mathfrak{w}$ in Section 3.3.2, this process will go through the weave until it reaches its right-hand side at the kth appearance of the lowest Coxeter generator. Figure 25 depicts examples of such faces in purple. Note that, as depicted on the right of the second row in that figure, the I–cycle might go through hexagonal vertices but shall always have the kth lowest Coxeter generator in $\mathfrak{w}_{0,n}$ at the two ends.

Second, for the two faces that involve the unique vertical edge, the associated absolute cycle in $\mathfrak{w}$ is the unique I–cycle that starts with the jth appearance of the lowest Coxeter generator at its boundary end (see Section 3.5.1) and has the other end at the unique trivalent vertex of $\mathfrak{w}$. Figure 25 depicts examples
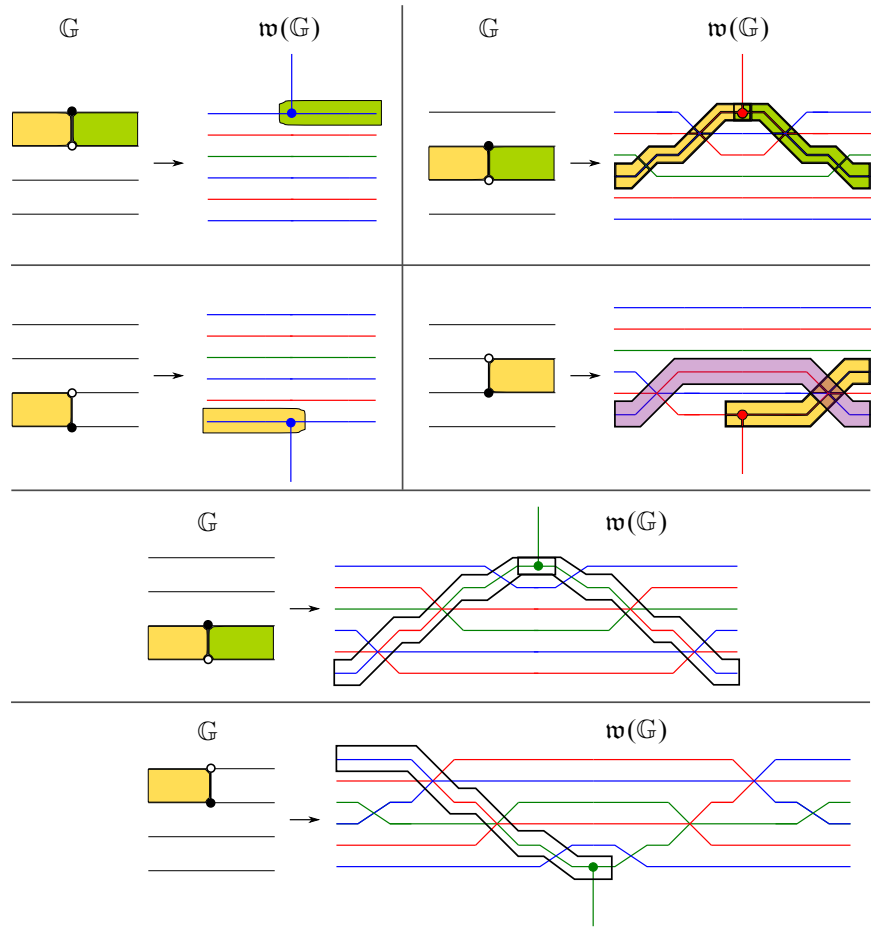
Figure 25: Several examples of faces in Type 2 elementary columns with $n = 4$ $\mathbb{G}$–strands, and their associated I–cycle in the weaves $\mathfrak{w}(\mathbb{G})$.

of such faces in yellow and green. Observe that, in general, these I–cycles will also go through hexagonal vertices but always have the jth lowest Coxeter generator at its boundary end.

3.5.3 Local representatives of naive absolute cycles in a Type 3 column

In a Type 3 column, the majority of faces are similar to a face in a Type 1 column. In the weave, their boundaries are represented by weave lines going across the weave as the unique long I–cycles with the correct boundary conditions. The only exceptional face in a Type 3 column is the face f which contains a lollipop, which we will discuss in detail in this subsection.

Let us first consider the case of a white lollipop attaching to the jth horizontal line on the right. For simplicity let us assume that the horizontal lines on the right are indexed by $1, 2, \ldots, n$ starting from the bottom. Following Section 3.5.1, the leftmost and rightmost ends of the cycle γ_f are determined by the Type 1 rules. Namely, given that the face f restricts to one gap on the left and two gaps on the right,

Type 3 column   cross-section of weave front Σ weave associated to this column
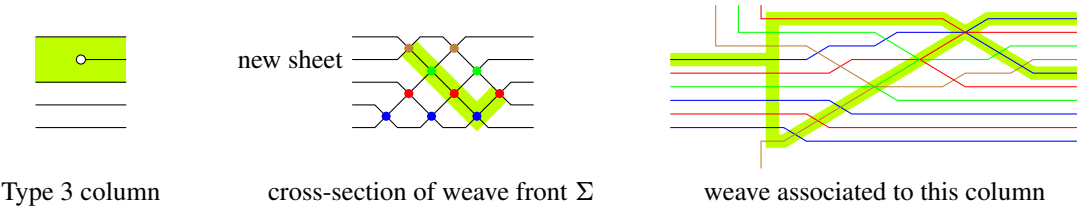
Figure 26: An elementary Type 3 column with a white lollipop (left) and its associated weave $\mathfrak{w}$ (right). Center: a vertical slice of the weave front highlighting the two directions of bifurcation that come up from the s_1–crossing (in blue) at the bottom.

the ends of the cycle γ_f in $\mathfrak{w}$ must be the unique I–cycle associated to those gaps. Thus, the cycle γ_f will start at a blue s_1–edge of the weave on the left and finish at two blue s_1–edges on the right. Now, in general, there does not exist a I–cycle (nor a Y–cycle) with these boundary conditions in $\mathfrak{w}$. This requires introducing a *bident*, as follows. Consider the middle slice of $\mathfrak{w}$ where all the newly emerged weave lines have become horizontal, ie the right boundary of the weave building block $\mathfrak{i}_j^w$ (Definition 3.16). Reading the weave lines from bottom to top at this slice yields an expression for the half-twist $w_{0,n} \in S_n$ (note that it is not $\mathfrak{w}_{0,n}$). Let us draw the weave slice as the positive braid $s_{[n-1,i]}^{-1} \mathfrak{w}_{0,n-1} s_{[n-1,n-i+1]}$ and mark the s_1–edge in $\mathfrak{w}_{0,n-1}$ that corresponds to the gap on the left within which the white lollipop emerges. Then, starting at this marked s_1–edge, we go along the upper-left and upper-right strands until we reach the highest (and last) possible crossing in each of the strands. These two crossings correspond to two weave lines on the right boundary of $\mathfrak{i}_j^w$. These two weave lines are said to be obtained from the (left) s_1–edge by a *bifurcation*.

Definition 3.25 A *bident* is a PL-embedding of a T–shape domain into the plane containing the weave such that on the left it coincides with an s_1–edge and on the right it coincides with the two crossings obtained by bifurcation on this s_1–edge.
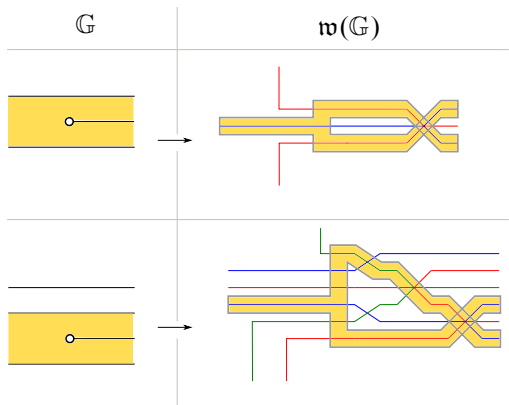
Figure 27: Left: examples of faces in Type 3 elementary columns with $n = 3, 4$ $\mathbb{G}$–strands. Right: the associated naive cycles, with the bidents, in the corresponding weaves $\mathfrak{w}(\mathbb{G})$.

a Type 3 column $\mathbb{G}$ associated weave $\mathfrak{w}(\mathbb{G})$
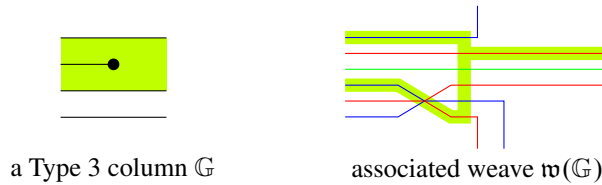
Figure 28: A case for the naive absolute cycle in a Type 3 column with a black lollipop. The face $f \in \mathbb{G}$ and its associated cycle $\gamma_f \subset \mathbb{R}^2$ are both highlighted in light green.

Finally, the case of an elementary column of Type 3 with a black lollipop is treated in exactly the same manner as for a white lollipop, with the roles vertically reversed; see Figure 28. For a face f containing a black lollipop in a Type 3 column, the boundary conditions being I–cycles on the weave and having a unique bident determine the cycle $\gamma_f \subset \mathbb{R}^2$ in the same manner as in the white lollipop case, except now the bident is left-pointing.

Due to the possible existence of bidents, it is hard to tell whether a naive absolute cycle has self-intersections (and hence it is not an embedded absolute cycle or $\mathbb{L}$–compressible) or not. It is easier if we can represent the naive absolute cycles as Y–cycles (Definition 3.4). Thus, we prove the following:

Proposition 3.26 *Let $\mathbb{G} \subset \mathbb{R}^2$ be a GP graph and $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ its associated weave. Then there exist a weave $\mathfrak{w}'$ and an equivalence $\mathfrak{w}' \sim \mathfrak{w}$ such that, under the isotopy[7] between $\Sigma(\mathfrak{w})$ and $\Sigma(\mathfrak{w}')$, the image of each naive absolute cycle on $L(\mathfrak{w})$ is homologous to a Y–cycle on $L(\mathfrak{w}')$.*

Proof For Type 1 and Type 2 elementary columns, the associated (pieces of) naive absolute cycles are already I–cycles, and hence Y–cycles. In particular, for $\mathbb{G}$ with no (internal) lollipops, we can take $\mathfrak{w}' = \mathfrak{w}$. It thus suffices to study the case of a Type 3 column, where a bident appears: it suffices to show that there exists a weave equivalence that allows us to replace a bident by a Y–cycle.

For a Type 3 column with a face f containing a white lollipop, this is done as follows. Consider the two weave lines to the right of the bident where the cycle γ_f propagates. By construction, these two weave lines intersect (to the right) at a unique hexagonal weave vertex of $\mathfrak{w}(\mathbb{G})$. In addition, the horizontal weave line entering from the left at this hexagonal vertex connects with an I–cycle to the s_1–edge on the left of the bident where γ_f starts. Therefore, we can consider the Y–cycle $\widetilde{\gamma}_f$ which starts with this s_1–edge at the left, propagates to the left (as an I–cycle) until the hexagonal vertex, and then contains a unique Y–vertex at the hexagonal vertex. Figure 29 depicts both cycles γ_f, at the left of the first row, and $\widetilde{\gamma}_f$, at the left of the second row, in the case of the Type 3 elementary column drawn in Figure 27, upper left. By considering the description of γ_f via vertical slices, it is readily seen that γ_f is homologous to $\widetilde{\gamma}_f$. In consequence, in the case of a white lollipop we can consider the same weave $\mathfrak{w}' = \mathfrak{w}$ and have the naive absolute cycle γ_f with a bident be homologous to the Y–cycle $\widetilde{\gamma}_f$.

---

[7]This isotopy naturally induces a Hamiltonian isotopy between $L(\mathfrak{w})$ and $L(\mathfrak{w}')$.
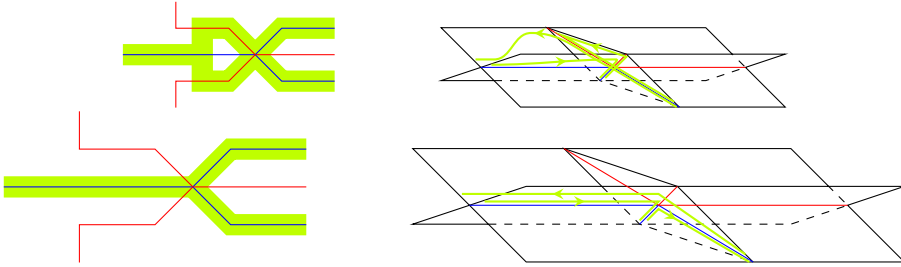
Figure 29: Two homologous cycles γ_f and $\widetilde{\gamma}_f$ depicted on the left. The upper-left cycle γ_f contains a bident and is not a Y–cycle; the lower-left cycle $\widetilde{\gamma}_f$ is a Y–cycle. The right-hand side of each row depicts these cycles in their spatial Legendrian fronts.

For a Type 3 column $\mathbb{G}$ with a black lollipop, the situation is similar, with the exception that a hexagonal vertex might not exist in $\mathfrak{w}(\mathbb{G})$ and thus the bident cannot readily be substituted by a Y–cycle. Nevertheless, we can insert two consecutive hexagonal vertices with a candy twist — Move I in Figure 12 — and then apply the same argument as above. □

3.6 Naive relative cycles in $L(\mathfrak{w}(\mathbb{G}))$

Let $L = L(\mathfrak{w}(\mathbb{G}))$ be the initial filling of the GP link $\Lambda = \Lambda(\mathbb{G})$. Section 3.5 constructed an explicit set of generators for a basis of $H_1(L)$ in terms of Y–cycles. In order to construct cluster $\mathcal{A}$–variables, we also need access to the lattice given by the relative homology group $H_1(L, \Lambda) = H_1(L, \partial L)$. Recall that, by Poincaré duality, there exists a nondegenerate pairing between the absolute homology group $H_1(L)$ and the relative homology group $H_1(L, \Lambda)$:

$$\langle \cdot, \cdot \rangle \colon H_1(L) \otimes H_1(L, \Lambda) \to \mathbb{Z}.$$

Let $\{\gamma_f\}$ be the basis of naive absolute cycles constructed in Section 3.5, where the index f runs over all faces of $\mathbb{G}$. Consider the Poincaré dual basis $\{\eta_f\}$ on $H_1(L, \Lambda)$.[8] In order to perform computations in the moduli stack of sheaves, we also want to describe the relative cycles in $\{\eta_f\}$ combinatorially in terms of the weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$. This is done according to the following discussion.

In general, given an N–weave $\mathfrak{w} \subset \mathbb{R}^2$, we can consider an (unoriented) curve $\kappa \subset \mathbb{R}^2$ that ends at unbounded regions in the complement of $\mathfrak{w} \subset \mathbb{R}^2$ and intersect weave lines of $\mathfrak{w}$ transversely and generically — in particular, away from the weave vertices. There are N natural ways to lift κ to the weave front $\Sigma(\mathfrak{w})$, which in turn correspond to N unoriented curves on L. In consequence, any subset of these N lifts, together with any orientation we choose for each of element of such a subset, defines a relative homology cycle $\eta \in H_1(L, \Lambda)$. Figure 30, left, depicts two possible oriented lifts of the (dashed) yellow curve κ drawn to its right.

---

[8] The dual of an absolute cycle γ_f is constructed from the entire basis of naive absolute cycles, not just γ_f.
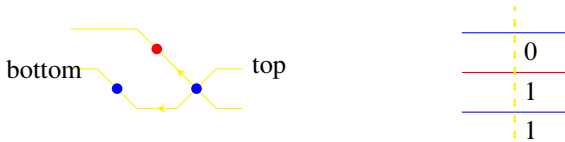
Figure 30: Left: in yellow, two lifts of the dashed curve κ depicted on the right. Right: a 3–weave with a labeled dashed curve κ; the labels 0, 1, 1 indicate the intersection number with each of the (pieces of cycles associated to) the weave lines.

Back to the case with $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$, where we have the basis of naive absolute cycles available, we can compute the intersection numbers between such relative cycles η and the naive absolute cycles. This leads to a tuple of integers $I(\eta) := (\langle \eta, \gamma_f \rangle)_{\text{faces } f}$. By the nondegeneracy of the Poincaré pairing $\langle \cdot, \cdot \rangle$ and the fact that $\{\gamma_f\}_{\text{faces } f}$ is a basis, the tuple of intersection numbers $I(\eta)$ uniquely determines the relative homology class of η. In fact, given that all naive absolute cycles that η intersects nontrivially must pass through weave lines, in order to describe the relative homology class of η it suffices to draw the unoriented curve $\kappa \subset \mathbb{R}^2$ and record the collection of the intersection number of its lift η with each of the weave lines. In order to distinguish such curves from weave lines, we will use dashed lines to depict such a curve κ.

Definition 3.27 A dashed curve $\kappa \subset \mathbb{R}^2$ as above, with the data of intersection numbers for each weave line it crosses, is called a *labeled dashed curve*.

Figure 30, right, depicts a labeled dashed curve κ. From a diagrammatic perspective, it is desirable to be manipulate labeled dashed curves in a weave diagram in the same manner that [Casals and Zaslow 2022] explained how to combinatorially manipulate absolute cycles. For that, we have depicted in Figure 31 the key moves on labeled dashed curves; these are all equivalences, in that these moves do not change the relative homology classes that the labeled dashed curves represent.

Finally, we can now diagrammatically describe a collection of labeled dashed curves that is a basis of the relative homology group $H_1(\Sigma, \Lambda)$, dual to the naive basis of $H_1(\Sigma)$ built in Section 3.5, as follows. First, for each face $f \subset \mathbb{G}$, we select a Type 1 elementary column inside of f. Second, consider the
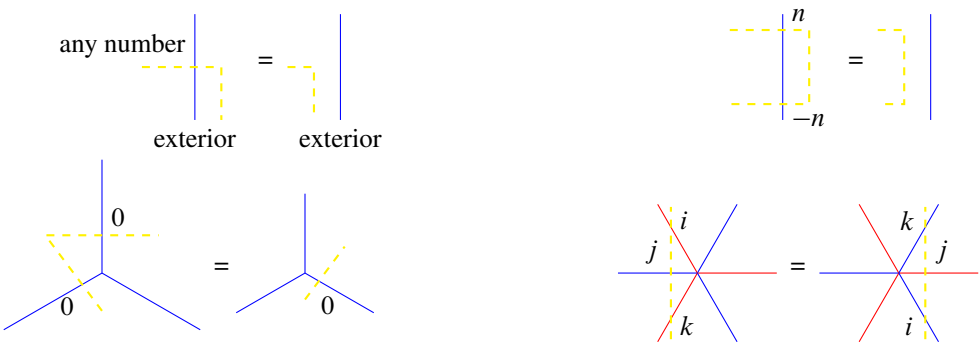
Figure 31: Four equivalence moves for labeled dashed curves.
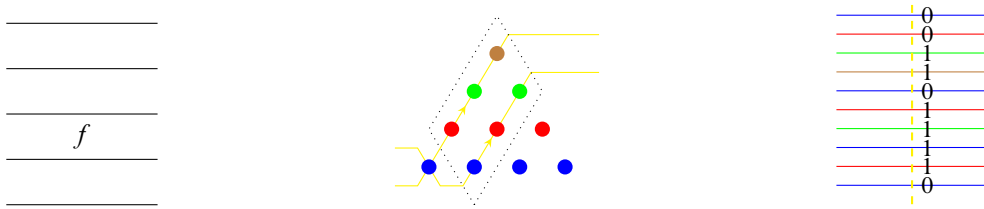
Figure 32: Left: a face $f \in \mathbb{G}$ chosen at an elementary Type 1 column. Center: a slice of the weave $\mathfrak{w}(\mathbb{G})$ with the rhomboid diamond associated to the s_1–crossing, in blue, for the face $f \in \mathbb{G}$. Right: the labeled dashed curve κ_f, in yellow, in the weave $\mathfrak{w}(\mathbb{G})$, with 1 in the weave lines for the crossings inside the diamond, and 0 otherwise.

piece of the weave $\mathfrak{w}(\mathbb{G})$ associated to this column — weave lines arranged according to $\mathfrak{w}_{0,n}$ — and draw a vertical dashed curve κ_f transverse to this piece of the weave. It now suffices to specify the correct labels encoding the intersection between the cycle for κ and the naive absolute cycle associated to this face $f \in \mathbb{G}$. For that, consider the braid given by slicing the weave front Σ along κ. Recall that there is a natural bijection between the appearances of the lowest Coxeter generator s_i in $\mathfrak{w}_{0,n}$ and the gaps in this Type 1 column (see Section 3.5.1). Locate the appearance of s_i that corresponds to a gap belonging to f and consider the set of crossings in this braid which are contained within the rhomboid diamond whose unique lowest vertex is at this appearance of s_i. Figure 32, center, draws an example of such a diamond for the face $f \subset \mathbb{G}$ depicted to its left. Then we label the curve κ, to a labeled curve κ_f, by assigning the intersection number 1 for all the weave lines in $\mathfrak{w}(\mathbb{G})$ which are associated to crossings in the braid *inside* the diamond, and by assigning the intersection number 0 for all the remaining weave lines. Figure 32, right, depicts the corresponding curve κ_f with its intersection labels for the face $f \subset \mathbb{G}$.

By construction, the intersection pairing between the labeled dashed curve κ_f and the naive absolute cycles is given by $\langle \eta_f, \gamma_f \rangle = 1$ and $\langle \eta_f, \gamma_g \rangle = 0$ for $g \neq f$. Thus, the collection of relative cycles $\{\eta_f\}$ associated to these particular labeled dashed curves κ_f are representatives of a dual naive basis of $H_1(L, \Lambda)$. We call this collection of relative cycles the *naive basis of relative cycles* of the relative homology group $H_1(L, \Lambda)$.

3.7 Initial absolute cycles and initial relative cycles in $L(\mathfrak{w}(\mathbb{G}))$

Sections 3.5 and 3.6 explain the construction of the naive basis of absolute cycles and the corresponding naive basis of relative cycles. The generators of these basis are not geometrically appropriate: despite being Y–cycles in the weave (or dual to them), they are often represented by *immersed* cycles and it is a priori unclear whether it is possible to mutate at them.[9] A key idea in this manuscript is the consideration and study of *sugar-free* hulls, as introduced in Section 2.3. In this subsection, these two parts, sugar-free hulls and the study of homology cycles compatible with the weave $\mathfrak{w}(\mathbb{G})$, converge: we show that it is

---

[9]In any sense of the word mutation: geometrically, through a Lagrangian surgery; diagrammatically, via a weave mutation; or cluster-theoretically, mutating at the naive vertex representing them in the naive quiver.
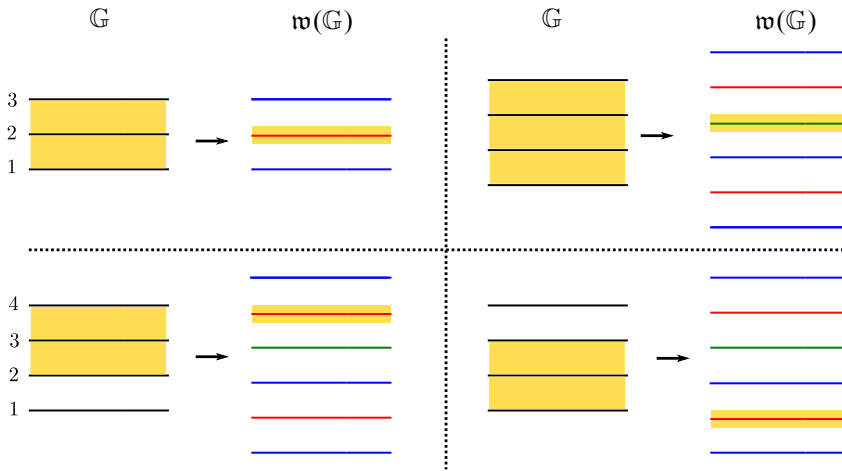
Figure 33: Associating a (piece of a) cycle in the weave for regions on a Type 1 column. The first row depicts the case where the region is the entire column for $n = 3$ and 4 strands. The second row depicts the remaining two cases for $n = 4$ strands.

possible to associate a Y–tree absolute cycle on $\mathfrak{w}(\mathbb{G})$ to every sugar-free hull of $\mathbb{G}$. In consequence, given that Y–trees are *embedded*, it will be possible to perform a weave mutation at every sugar-free hull of $\mathbb{G}$. This leads to the notion of *initial basis*, which eventually give rise to the initial seeds for our cluster structures.

In the study of sugar-free hulls, we must consider cycles which are associated to regions of $\mathbb{G}$ — namely the sugar-free hulls — and not just faces $f \subset \mathbb{G}$. The simplest case is that of a region in an elementary column of Type 1, which is considered in the following simple lemma, where we use the weave $\mathfrak{n}(\mathtt{w}_{0,n})$ introduced in Definition 3.11:

Lemma 3.28 *Let $\mathbb{G}$ be a GP graph and $C \subset \mathbb{G}$ a Type 1 elementary column. Consider the region $R \subset C$ given by the union of k consecutive gaps in C. Then the boundary of ∂R is homologous to the lift of a unique weave line on the kth level.*

Proof The boundary of a single gap has two connected components, and each of them is a deformation retract of a sheet of the spatial wavefront $\Sigma(C)$. For a single gap, the two sheets associated with its boundary intersect at a unique weave line at the bottom level. For a union R of k consecutive gaps, the two sheets associated with ∂R intersect at the kth level. □

Figure 33 depicts four cases illustrating how to associate a cycle on a weave line for a region on a Type 1 column. The case of arbitrary strands can be readily imagined by examining these few cases.

Proposition 3.26 showed that it is possible, up to possibly performing a weave equivalence, to represent the naive absolute cycles with Y–cycles. But these are typically immersed: it is not always possible,
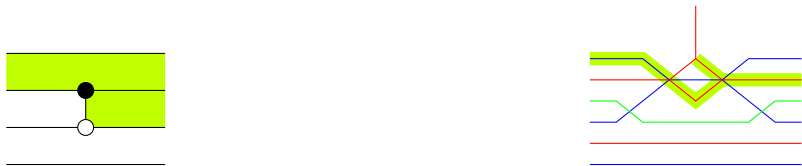
Figure 34: Left: Type 2 column with region R highlighted in light green. Right: the associated local weave and the Y–tree, in the shape of a tripod.

in general, to find embedded Y–cycles representing these homology classes. Now, the following result, which we refer to as the Y–representability lemma, shows that it *is* possible to represent the boundary cycle ∂R by a Y–tree if the region R is sugar-free:

Lemma 3.29 (Y–representability lemma) *Let $\mathbb{G}$ be a GP graph and $R \subset \mathbb{G}$ a sugar-free region. Then the boundary cycle ∂R is homologous to a Y–tree on $\mathfrak{w}(\mathbb{G})$, up possibly performing a weave equivalence that adds $\mathfrak{n}_k^{\uparrow}(\mathbb{w}_{n,0})\mathfrak{n}_k^{\uparrow}(\mathbb{w}_{n,0})^{\mathrm{op}}$ and $\mathfrak{n}_k^{\downarrow}(\mathbb{w}_{n,0})\mathfrak{n}_k^{\downarrow}(\mathbb{w}_{n,0})^{\mathrm{op}}$ to $\mathfrak{w}(\mathbb{G})$.*

Proof Let C be an elementary column of $\mathbb{G}$. By Lemma 2.4, the intersection $R \cap C$ has at most one connected component. Therefore, if R intersects C nontrivially, R must be a union of consecutive gaps in the column C. By Lemma 3.28, for a Type 1 or Type 3 elementary column C, we can represent the boundary $\partial(R \cap C)$ by a single I–cycle weave line going from left to right; note that the weave line color may change within a Type 3 column. It thus remains to treat the cases of elementary columns of Type 2. By Lemma 2.3 applied to a Type 2 column, we conclude that the following four cases — in correspondence with the four staircase patterns — are to be analyzed:

First, let us consider the staircase pattern which is second from the left. The corresponding local weave pattern is $\mathfrak{c}_k^{\uparrow}(\mathbb{w}_{0,n})$, as introduced in Definition 3.13. In the construction of this weave pattern, we first bring the kth strand in the bottom level upward, using the weave pattern $\mathfrak{n}_k^{\uparrow}(\mathbb{w}_{0,n})$, subsequently insert a trivalent weave vertex at the top strand, and then insert the weave pattern $\mathfrak{n}_k^{\uparrow}(\mathbb{w}_{0,n})^{\mathrm{op}}$. Figure 34 depicts a case with four horizontal lines and $k = 2$. (See also Figures 17 and 18.) Since the kth horizontal line is the bottom boundary of ∂R, there must be a unique hexavalent weave vertex in the $\mathfrak{n}_k^{\uparrow}(\mathbb{w}_{0,n})^{\mathrm{op}}$ that connects to the weave line corresponding to the union of all gaps in R at the right boundary. Therefore,
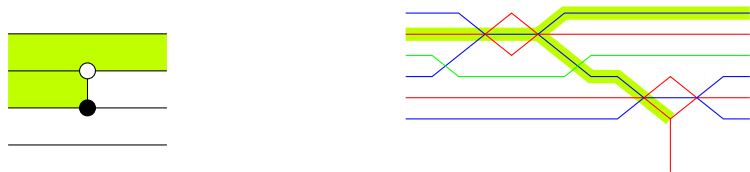
Figure 35: Example of the first staircase.

we can create a Y–tree in this region, with the required boundary conditions, by inserting a tripod leg that connects to the trivalent weave vertex (at the top) and continues to the left towards whichever weave line is required by the boundary condition of R. This resulting Y–tree, in the shape of a tripod, then represents ∂R locally, as desired. Figure 34, right, depicts this Y–tree, highlighted in light green, for the region R on the left, also drawn in the same color. This concludes the second case among the four staircase patterns; the third case, which also contains the region R to the right of the crossing, can be resolved analogously.

Next let us study the leftmost of the four staircase patterns. In this case, the chosen weave pattern $\mathfrak{c}_k^{\downarrow}(\mathfrak{w}_{0,n})$, as assigned in Section 3.3, does not already have a hexavalent weave vertex that meets our need. Nevertheless, we can create it by concatenating the local weave pieces $\mathfrak{n}_k^{\uparrow}(\mathfrak{w}_{0,n})$ and $\mathfrak{n}_k^{\uparrow}(\mathfrak{w}_{0,n})^{\mathrm{op}}$ on the left first. Note that this can be achieved by inserting a series of Moves I and V, and thus the equivalence class of the weave remains the same. Now, inside of the weave piece $\mathfrak{n}_k^{\uparrow}(\mathfrak{w}_{0,n})^{\mathrm{op}}$, there exists a unique hexavalent weave vertex that connects to both the weave line representative of ∂R, to the left, and the trivalent weave vertex in $\mathfrak{c}_k^{\downarrow}(\mathfrak{w}_{0,n})$, to the right. The Y–tree, again in a tripod shape, represents ∂R locally, as desired. Figure 35 depicts an example of such a tripod with $n = 4$ and $k = 2$. An analogous argument also resolves the case of the rightmost staircase pattern. Finally, by combining the local pictures for all three types of column, we conclude that there exists a representative for ∂R which is a Y–tree. □

The Y–representability lemma allows us to introduce the following definition:

Definition 3.30 Let $\mathbb{G}$ be a GP graph. The set of *initial absolute cycles* $\mathfrak{S}(\mathbb{G})$ is the set of all Y–trees on $L(\mathfrak{w}(\mathbb{G}))$ which are associated to the (nonempty) sugar-free hulls in $\mathbb{G}$.

The inclusion relation between sugar-free hulls naturally puts a partial order on the set $\mathfrak{S}(\mathbb{G})$: by definition, $\partial R \leq \partial R'$ if $R \subset R'$ as sugar-free hulls. Also, given a face $f \subset \mathbb{G}$, we note that a face $g \in \mathbb{S}_f$ in its sugar-free hall, must satisfy $\mathbb{S}_g \subset \mathbb{S}_f$ and hence $\partial \mathbb{S}_g \leq \partial \mathbb{S}_f$. Finally, note that multiple faces in $\mathbb{G}$ can share the same sugar-free hull, and there may exist faces with an empty sugar-free hull. Thus, in general, the number of sugar-free hulls may be smaller than the number of faces in $\mathbb{G}$. Nevertheless, we can prove that the set of initial absolute cycles $\mathfrak{S}(\mathbb{G})$ is always linearly independent.

Proposition 3.31 *Let $\mathbb{G}$ be a GP graph and let $\mathfrak{S}(\mathbb{G})$ be its set of initial absolute cycles. Then $\mathfrak{S}(\mathbb{G})$ is a linearly independent subset of $H_1(L(\mathfrak{w}(\mathbb{G})))$. In addition, it is possible to add naive absolute cycles to $\mathfrak{S}(\mathbb{G})$ to complete $\mathfrak{S}(\mathbb{G})$ into a basis of $H_1(L(\mathfrak{w}(\mathbb{G})))$.*

Proof Let $L = L(\mathfrak{w}(\mathbb{G}))$ be the initial filling. Consider the naive basis $\{\gamma_f\}_{\text{faces } f}$ of $H_1(L)$. It suffices to show that we can replace $|\mathfrak{S}(\mathbb{G})|$ many naive basis elements with elements in $\mathfrak{S}(\mathbb{G})$ while maintaining the spanning property of the set. This is can be done by using the partial order on $\mathfrak{S}(\mathbb{G})$ as follows.

Let us start with the minimal elements in $\mathfrak{S}(\mathbb{G})$ and work our way up, replacing the appropriate naive basis elements in $\{\gamma_f\}$ with elements in $\mathfrak{S}(\mathbb{G})$. At each turn, we select a face $f \subset \mathbb{G}$ whose sugar-free

hull $\mathbb{S}_f$ defines a Y–tree $\partial\mathbb{S}_f \in \mathfrak{S}(\mathbb{G})$ and then replace the naive absolute cycle γ_f with the Y–tree $\partial\mathbb{S}_f$. Note that it is always possible to find such a face $f \subset \mathbb{G}$ in this process: if no such face f were available, any faces within the sugar-free hull would not have this particular sugar-free region as their sugar-free hull, which is tautologically absurd.

In the process of implementing each of these replacements, we must argue that the resulting set still spans the homology group $H_1(L)$. For that we observe that, if a replacement of γ_g by $\partial\mathbb{S}_g$ is done before the replacement of γ_f by $\mathbb{S}_f$, the sugar-free hull $\mathbb{S}_g$ must not contain the face f inside. This follows from the fact that $f \in \mathbb{S}_g$ implies $\mathbb{S}_f \subset \mathbb{S}_g$. Therefore, we can use the set $\{\partial\mathbb{S}_g \mid g \in \mathbb{S}_f\}$, which is already in the basis set due to the partial order, and $\partial\mathbb{S}_f$ to recover $\gamma_f = \partial f$. This shows that the resulting set after the replacement of γ_f by $\partial\mathbb{S}_f$ as above is still a basis for $H_1(L)$. □

The choice of completion of $\mathfrak{S}(\mathbb{G})$ to a basis in $H_1(L(\mathfrak{w}(\mathbb{G})))$ is a neat instance of the natural appearance of quasicluster structures in (symplectic) geometry. There is no particular canonical manner by which we can typically choose a basis for this complement, but, as we shall explain, the different choices all lead to the same cluster structure *up to* monomials in the frozen variables, ie a quasicluster structure. The dualization process, via the Poincaré pairing, requires a choice of basis. Therefore, there is no canonical choice of *initial relative cycles* associated to the set $\mathfrak{S}(\mathbb{G})$ unless the latter spans $H_1(L(\mathfrak{w}(\mathbb{G})))$. In a general situation, we can at least consider the following concept:

Definition 3.32 Let $\mathbb{G}$ be a GP graph, $L = L(\mathfrak{w}(\mathbb{G}))$, $\Lambda = \Lambda(\mathbb{G})$ and $\mathfrak{S}(\mathbb{G})$ its set of initial absolute cycles on L. Let $\mathfrak{B}$ be a basis of $H_1(L)$ which is obtained by adding naive absolute cycles to the set $\mathfrak{S}(\mathbb{G})$. Then the *initial relative cycles* associated to $\mathfrak{B}$ is the collection $\mathfrak{B}^\vee$ of linear combinations of naive relative cycles whose relative homology classes form a basis of $H_1(L, \Lambda)$ dual to the basis $\mathfrak{B}$ of $H_1(L)$.

We emphasize that the basis $\mathfrak{B}^\vee$ depends not only on $\mathfrak{S}(\mathbb{G})$, but also on the chosen basis completion $\mathfrak{B}$. Note that a few simple choices of $\mathfrak{B}$ are available as a result of the replacement construction in the proof of Proposition 3.31. Namely, we start with the naive absolute basis $\{\gamma_f\}$, and then swap some of the basis elements γ_f with their corresponding initial absolute cycles $\partial\mathbb{S}_f$. In the case that there are multiple faces sharing the same sugar-free hull, only one of the naive absolute cycles gets replaced, and the rest remain in the basis, which will become frozen basis elements.

Furthermore, if the basis $\mathfrak{B}$ is chosen via such a basis replacement process, then the corresponding dual relative cycle basis $\mathfrak{B}^\vee$ can be described with respect to the partial order on sugar-free hulls as well. Indeed, suppose that we have an equality $\mathbb{S}_f = \mathbb{S}_g = \cdots$ of sugar-free hulls for some faces $f, g \subset \mathbb{G}$, among others, and γ_f was chosen to be replaced by $\partial\mathbb{S}_f$ in the replacement process. Then the naive relative cycle η_g needs to be replaced by $\eta_g - \eta_f$ for each g with $\mathbb{S}_g = \mathbb{S}_f$. Similarly, η_f would need to be replaced by $\eta_f + N$, where the N summand is a linear combination of the naive relative cycles η_h associated to the chosen faces h with $\mathbb{S}_f \subsetneq \mathbb{S}_h$, so that the pairing of $\eta_f + N$ with $\partial\mathbb{S}_h$ vanishes for all such h.

3.8 The naive quiver of a GP graph

Let us start by emphasizing that the quiver of the initial seed for the cluster structure we construct is *not* always the dual quiver of the GP graph $\mathbb{G}$. Nevertheless, that dual quiver is useful in order to construct the actual quiver of the initial seed because it can be used to compute the intersection form on the initial filling, which is needed to define the initial quiver. Let us provide the details.

Following [Gross et al. 2018], in order to define a cluster structure, we first fix an integer lattice with a skew-symmetric form on it. In the context of a GP graph $\mathbb{G}$, the integer lattice is $H_1(L(\mathfrak{w}(\mathbb{G})))$, and a natural skew-symmetric form on it is given by the intersection pairing between absolute homology classes. By using the *naive* basis $\{\gamma_f\}$ of naive absolute cycles and the GP graph $\mathbb{G}$, we can describe the intersection pairing form $\{\cdot,\cdot\}$ combinatorially using a quiver.

Definition 3.33 Let $\mathbb{G} \subset \mathbb{R}^2$ be a GP graph. The *naive quiver* $Q_0(\mathbb{G})$, or *dual quiver*, associated to $\mathbb{G}$ is the quiver constructed as follows:

(1) A quiver vertex is associated to each face $f \subset \mathbb{G}$.

(2) For every bipartite edge in $\mathbb{G}$, we draw an arrow according to $\bullet\!-\!\!\!\dashrightarrow\!\!\bullet$.

(3) For each pair of quiver vertices, sum up the arrows between them.

Note that in step (3) there might be cancellations.

Lemma 3.34 [Goncharov and Kenyon 2013, Definition 8.2 and Proposition 8.3] *Let ϵ_{fg} be the exchange matrix of the quiver $Q_0(\mathbb{G})$. Then the intersection pairing between γ_f and γ_g is given by $\{\gamma_f, \gamma_g\} = \epsilon_{fg}$.*

Since $\{\gamma_f\}_{\text{faces } f}$ is a basis of $H_1(L(\mathfrak{w}(\mathbb{G})))$, Lemma 3.34 uniquely determines the intersection skew-symmetric form. This intersection form, ie the quiver $Q_0(\mathbb{G})$, is then used to compute the correct quiver $Q(\mathbb{G})$ for the initial seed. The unfrozen vertices of the correct initial quiver $Q(\mathbb{G})$ will be indexed by $\mathfrak{S}(\mathbb{G})$, the set of initial absolute cycles, and the remaining frozen vertices are determined by the choice of completion of $\mathfrak{S}(\mathbb{G})$ to a basis of $H_1(L)$. Since we will elaborate more on this in Section 4, we conclude this discussion for now and revisit $Q(\mathbb{G})$ then.

Remark 3.35 In the case of a plabic fence $\mathbb{G}$, all naive absolute cycles are I–cycles and thus they also are initial absolute cycles. Thus, for a plabic fence, $Q(\mathbb{G})$ coincides with $Q_0(\mathbb{G})$.

3.9 Bases and homology lattices in the presence of marked points

The construction of the cluster structures in Theorem 1.1, and the definition of the moduli space $\mathfrak{M}(\Lambda, T)$, in general require an additional piece of data: a set T of marked points on $\Lambda(\mathbb{G})$.

Definition 3.36 Let $\mathbb{G}$ be a GP graph and let $\Lambda = \Lambda(\mathbb{G})$ be its GP link. A set of *marked points* $T \subset \Lambda$ is a subset of distinct points in Λ, where we require that there is at least one marked point on each link component of Λ and, without loss of generality, the set T is disjoint from all crossings and all cusps in the front $\mathfrak{f}(\mathbb{G})$ of Λ.

All prior statements in Section 3 remain unchanged by the addition of marked points, as they do not affect the associated weaves or the Hamiltonian isotopy class of exact Lagrangian fillings. Therefore, we can still consider the initial embedded exact filling $L = L(\mathfrak{w}(\mathbb{G}))$ of the GP link Λ. As before, we select the collection of initial absolute cycles $\mathfrak{S}(\mathbb{G})$ associated with sugar-free hulls, and they form a linearly independent subset of $H_1(L)$. The addition of marked points affects only the cluster-theoretic constructions: we need to replace the lattice of absolute homology $H_1(L)$ by the lattice of relative homology $H_1(L, T)$. The natural inclusion $H_1(L) \subset H_1(L, T)$, induced by the inclusions $T \subset \Lambda = \partial L \subset L$, allows us to include the initial absolute cycles $\mathfrak{S}(\mathbb{G})$ as a linearly independent subset of $H_1(L, T)$. The only difference is that, in order to fix a cluster structure, we must expand $\mathfrak{S}(\mathbb{G})$ further to a basis $\mathfrak{B}$ of $H_1(L, T)$. This expansion can be done in two steps: we first expand $\mathfrak{S}(\mathbb{G})$ to a basis of $H_1(L)$, as done via the replacement process in Section 3.7, and then expand this basis of $H_1(L)$ to a basis of $H_1(L, T)$.

As was the case for $H_1(L)$ and its dual $H_1(L, \Lambda)$, we shall need a dual space of $H_1(L, T)$ together with a basis dual to a chosen basis $\mathfrak{B}$ of $H_1(L, T)$. In fact, there is a natural intersection pairing

$$\langle \cdot, \cdot \rangle \colon H_1(L, T) \otimes H_1(L \setminus T, \Lambda \setminus T) \to \mathbb{Z}$$

obtained by algebraically counting geometric intersections of relative cycles in generic position. In the same manner that Poincaré duality was used in Section 3.7, a duality also exists in the setting with marked points. We record the precise statement in the following:

Proposition 3.37 *Let L be a connected smooth surface with boundary $\Lambda = \partial L$, and $i \colon T \to \Lambda$ an inclusion of a set of marked points with $\pi_0(i)$ surjective. Then*

$$\mathrm{rk}(H_1(L, T)) = \mathrm{rk}(H_1(L \setminus T, \Lambda \setminus T)),$$

and the intersection pairing $\langle \cdot, \cdot \rangle$ is nondegenerate.

It is possible to consider intermediate lattices M and N in between the lattices discussed above. Namely, we can consider sublattices N of $H_1(L, T)$ which include $H_1(L)$, and dually quotients M of $H_1(L \setminus T, \Lambda \setminus T)$, as in the following diagram, where all horizontal arrows are dual lattices:

$$
\begin{array}{ccc}
H_1(L, T) & \longleftrightarrow & H_1(L \setminus T, \Lambda \setminus T) \\
\uparrow & & \downarrow \\
N & \longleftrightarrow & M \\
\uparrow & & \downarrow \\
H_1(L) & \longleftrightarrow & H_1(L, \Lambda)
\end{array}
$$

4   Construction of quasicluster structures on sheaf moduli

In this section we develop the necessary results to study the geometry of the moduli stack $\mathfrak{M}(\Lambda, T)$ associated to $\Lambda = \Lambda(\mathbb{G})$ and prove Theorem 1.1. In particular, we introduce microlocal merodromies in Section 4.6 which, as we will prove, become the cluster $\mathscr{A}$–variables. The construction of the cluster structures is obtained purely by symplectic geometric means, using the results for Legendrian weaves from Section 3 above and [Casals and Zaslow 2022] and the microlocal theory of sheaves [Guillermou et al. 2012; Kashiwara and Schapira 1990; Shende et al. 2017]. Let us review what we have developed in Sections 2 and 3 thus far. Given a GP graph $\mathbb{G}$, we constructed the following list of objects:

(i)   A Legendrian link $\Lambda = \Lambda(\mathbb{G})$, which is a (-1)–closure of a positive braid $\beta(\mathbb{G})$.

(ii)   An exact Lagrangian filling $L = L(\mathfrak{w})$ of Λ, called the *initial filling*. This exact Lagrangian filling L is obtained as the Lagrangian projection of the Legendrian lift associated with the spatial front defined by the *initial weave* $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$.

(iii)   A collection of *initial absolute cycles* $\mathfrak{S}(\mathbb{G})$, which form an $\mathbb{L}$–compressing system for L and can be described by Y–trees on $\mathfrak{w}$.

(iv)   A skew-symmetric intersection pairing on the lattice $H_1(L)$. This intersection pairing can be computed directly from the GP graph $\mathbb{G}$.

By specifying an additional generic set of *marked points* $T \subset \Lambda$ with at least one marked point per component, we also obtain the lattice $H_1(L, T)$, which contains $H_1(L)$ and hence the linearly independent subset $\mathfrak{S}(\mathbb{G})$. The skew-symmetric pairing on $H_1(L)$ extends naturally to a skew-symmetric pairing on $H_1(L, T)$. By Poincaré duality, we can identify the dual lattice of $H_1(L, T)$ with the relative homology $H_1(L \setminus T, \Lambda \setminus T)$. Any completion of $\mathfrak{S}(\mathbb{G})$ to a basis $\mathfrak{B}$ of $H_1(L, T)$ gives rise to a unique dual basis $\mathfrak{B}^\vee$ of $H_1(L \setminus T, \Lambda \setminus T)$.

The outline for this section is as follows. First, we give working definitions of the moduli space $\mathfrak{M}(\Lambda, T)$, which allows us to draw connections to Lie-theoretical moduli spaces and also deduce the factoriality of its ring of regular functions $\mathbb{O}(\mathfrak{M}(\Lambda, T))$. Next, on the moduli space $\mathfrak{M}(\Lambda, T)$, we construct a new family of rational functions called *microlocal merodromies*, which are associated with relative cycles in $H_1(L \setminus T, \Lambda \setminus T)$. Although the definition of microlocal merodromies depends on the initial filling L, we show that, for elements in the dual basis $\mathfrak{B}^\vee$, their microlocal merodromies actually extend to $\mathbb{C}$–valued regular functions on the entire moduli space $\mathfrak{M}(\Lambda, T)$. Moreover, we prove that, within these special microlocal merodromies, those dual to $\mathfrak{S}(\mathbb{G})$ can be mutated according to the cluster $\mathscr{A}$–mutation formula as the initial weave $\mathfrak{w}$ undergoes weave mutation, corresponding to a Lagrangian disk surgery on $L(\mathfrak{w})$. Then we show that the codimension 2 argument in cluster varieties can be applied by studying immersed Lagrangian fillings represented by nonfree weaves. These results together with [Berenstein et al. 2005] allow us to conclude the existence of a cluster $\mathscr{A}$–structure on $\mathfrak{M}(\Lambda, T)$, where the initial and adjacent seeds are constructed via the Lagrangian filling $L(\mathbb{G})$, its Lagrangian surgeries and the associated microlocal merodromies.

4.1 Descriptions of sheaves with singular support on the Legendrian $\Lambda(\mathbb{G})$

Let $\Lambda \subset (\mathbb{R}^3 \xi_{st})$ be a Legendrian link and $T \subset \Lambda$ a set of marked points, and consider the moduli stacks $\mathcal{M}_1(\Lambda)$ and $\mathfrak{M}(\Lambda, T)$ discussed in Section 2.7. These stacks classify (complexes of) constructible sheaves on $\mathbb{R}^2$ with a singular support condition. In this subsection, we provide Lie-theoretical descriptions for $\mathcal{M}_1(\Lambda)$ and $\mathfrak{M}(\Lambda, T)$ which are suited for our computations, using [Kashiwara and Schapira 1990] and closely following [Shende et al. 2017, Sections 3.3 and 5].[10] These are more combinatorial presentations of these stacks, as the constructible and microlocal aspects of the original definition are translated into explicit quiver representations satisfying certain conditions.

Given a cooriented front projection $\pi_F(\Lambda)$, consider the following quiver $Q_F(\Lambda)$:

- A vertex of $Q_F(\Lambda)$ is placed at each connected component of $\mathbb{R}^2 \setminus \pi_F(\Lambda)$,
- For each (1–dimensional) connected component of $\pi_F(\Lambda) \setminus S_0$, where S_0 denotes the set of crossings and cusps in $\pi_F(\Lambda)$, draw an arrow connecting the two vertices associated to the two adjacent 2–dimensional cells (that contain that stratum in their closure). The direction of the arrow is opposite to the coorientation of the front $\pi_F(\Lambda)$.

The following is then proven in [Shende et al. 2017, Section 3]:

Proposition 4.1 *Let* $\Lambda \subset (\mathbb{R}^3, \xi)$ *be a Legendrian and* $\pi_F(\Lambda) \subset \mathbb{R}^2$ *a front, with a binary Maslov potential, such that* $(\mathbb{R}^2, \pi_F(\Lambda))$ *is a regular stratification. Consider the stack* $\mathcal{M}(Q_F(\Lambda))$ *classifying linear representations of the quiver* $Q_F(\Lambda)$ *that satisfy the following conditions:*

(1) *The vector space associated with the unbounded region in* $\mathbb{R}^2$ *is* 0.

(2) *Any two vector spaces associated with neighboring vertices differ in dimension by* 1.

(3) *At each cusp, the composition depicted in Figure 36, left, is the identity map.*

(4) *At each crossing, the four linear maps involved form a commuting square which is exact, as precised in Figure 36, right.*

Then the stack $\mathcal{M}_1(\Lambda)$ *is isomorphic to* $\mathcal{M}(Q_F(\Lambda))$.

Proposition 4.1 describes $\mathcal{M}_1(\Lambda)$. Now we gear towards the decorated moduli $\mathfrak{M}(\Lambda, T)$. First, we need a description of microlocal monodromy in terms of these quiver representations, which is provided in [Shende et al. 2017, Section 5], and we briefly summarize as follows. Let S be the set of singular points (crossings and cusps) in $\pi_F(\Lambda)$ and note that each connected components of $\pi_F(\Lambda) \setminus S$ is associated with a 1–dimensional kernel or a 1–dimensional cokernel. These kernels and cokernels can be glued together along strands of Λ using the identity condition at cusps and the exactness condition at crossings, as follows:

---

[10]An expert in the results of [Kashiwara and Schapira 1990; Shende et al. 2017] might be able to quickly move forward to Section 4.2.

Figure 36: Identity condition at cusps and exactness condition at crossings.

- At a cusp, as in Figure 36, left, the condition $g \circ f = \mathrm{id}_V$ forces the composition $\ker g \hookrightarrow W \twoheadrightarrow \mathrm{coker}\, f$ to be an isomorphism. By definition, we glue $\ker g$ and $\mathrm{coker}\, f$ using this isomorphism.

- At a crossing, as in Figure 36, right, there are three cases, depending on the injectivity or surjectivity of the four maps: the four maps can be all injective, all surjective, or two injective with two surjective. In each of the three cases, we have the following isomorphisms from the exactness condition:

(4-1)
$$\mathrm{coker}\, f_{sw} \hookrightarrow \frac{V_n}{V_s} \twoheadrightarrow \mathrm{coker}\, f_{en},$$

$$\mathrm{coker}\, f_{wn} \twoheadleftarrow \frac{V_n}{V_s} \hookleftarrow \mathrm{coker}\, f_{se},$$

(4-2)
$$\ker f_{sw} \hookrightarrow \ker(f_{wn} \circ f_{sw}) = \ker(f_{en} \circ f_{se}) \twoheadrightarrow \ker f_{en},$$

$$\ker f_{wn} \twoheadleftarrow \ker(f_{wn} \circ f_{sw}) = \ker(f_{en} \circ f_{se}) \hookleftarrow \ker f_{se},$$

(4-3)
$$\mathrm{coker}\, f_{sw} \xrightarrow{f_{wn}} \mathrm{coker}\, f_{en},$$

$$\ker f_{wn} \xleftarrow{f_{sw}} \ker f_{se}.$$

The result of gluing these 1–dimensional vector spaces is a rank-1 local system Φ on Λ. In fact, it coincides with the microlocal monodromy functor; see [Shende et al. 2017, Section 5.1] for more details. Given the set T of marked points on Λ, with at least one marked point per link component, $\Lambda \setminus T$ is a collection of open intervals. Thus, along each such open interval I, we can trivialize the rank-1 local system Φ by specifying an isomorphism $\phi_I : I \times \mathbb{C} \xrightarrow{\cong} \Phi|_I$. By definition, a collection of such maps $\{\phi_I\}$ are said to be a *framing* for the local system Φ.

In conclusion, a point in the decorated moduli space $\mathfrak{M}(\Lambda, T)$, as defined in Section 2.7.3, is a point in $\mathcal{M}_1(\Lambda)$, which is combinatorialized via Proposition 4.1, together with a framing for the local system Φ, ie a trivialization of the (trivial) local system $\Phi|_{\Lambda \setminus T}$. Here two framings are considered *equivalent* if they differ by a global scaling $\mathbb{C}^\times$ factor, and thus $\dim \mathfrak{M}(\Lambda, T) = \dim \mathcal{M}_1(\Lambda) + |T| - 1$.

4.1.1 Description for a GP graph $\mathbb{G}$ In the case that $\Lambda = \Lambda(\mathbb{G})$ comes from a GP graph $\mathbb{G}$, Section 2.4 provides a specific front $\mathfrak{f}(\mathbb{G}) \subset \mathbb{R}^2$. For this front, the description from Proposition 4.1 can be translated
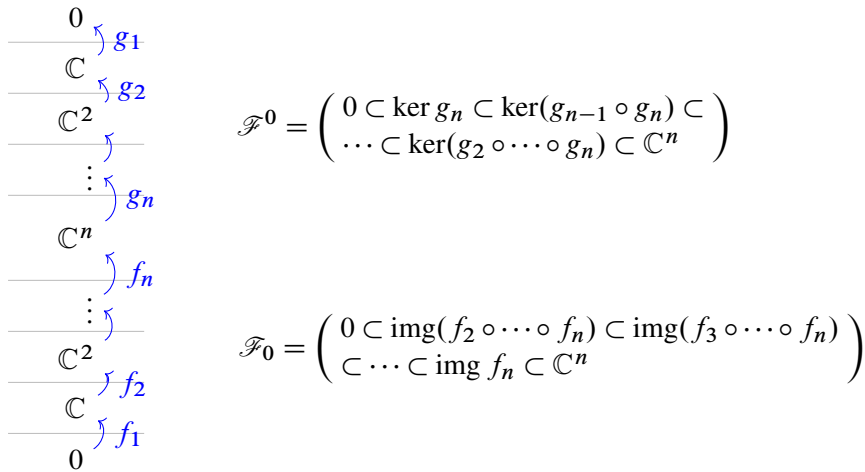
Figure 37: The pair of flags associated to a Type 1 column.

in terms of configurations of flags, as follows. The front projection $\pi_F(\Lambda) = \mathfrak{f}(\mathbb{G})$ can be sliced into the three types of elementary columns.

For a Type 1 column, there are n strands in the bottom region, with Maslov potential 0, and n strands in the top region, with Maslov potential 1. By Proposition 4.1, the vector space associated with the central region must be $\mathbb{C}^n$. From the quiver representation data, we can construct the following pair of flags in $\mathbb{C}^n$. Since all of the linear maps in the bottom region are injective, their images in the middle $\mathbb{C}^n$ naturally form a first flag. Similarly, since all of the linear maps in the top region are surjective, their kernels in the middle vector space $\mathbb{C}^n$ form a second flag. See Figure 37 for a depiction of the front in Type 1 and its associated pair of flags. We adopt the convention of indexing flags from the bottom region with a subscript, and indexing flags from the top region with a superscript, so as to distinguish between them.

Before discussing Type 2 and 3 columns, we recall that the relative positions relations between two flags in $\mathbb{C}^n$ are classified by elements of the symmetric group S_n, which is a Coxeter group with Coxeter generators $\{s_i\}_{i=1}^{n-1}$. For two flags $\mathscr{F} = (0 \subset \mathscr{F}_1 \subset \cdots \subset \mathbb{C}^n)$ and $\mathscr{F}' = (0 \subset \mathscr{F}_1' \subset \cdots \subset \mathbb{C}^n)$, we write

- $\mathscr{F} \overset{s_i}{\sim} \mathscr{F}'$ if $F_i \neq F_i'$ but $F_j = F_j'$ for all $j \neq i$;

- $\mathscr{F} \overset{w}{\sim} \mathscr{F}'$ if there exists a sequence of flags $\mathscr{G}_0, \mathscr{G}_1, \ldots, \mathscr{G}_l$ such that

$$\mathscr{F} = \mathscr{G}_0 \overset{s_{i_1}}{\sim} \mathscr{G}_1 \overset{s_{i_2}}{\sim} \mathscr{G}_2 \overset{s_{i_3}}{\sim} \cdots \overset{s_{i_l}}{\sim} \mathscr{G}_l = \mathscr{F}'$$

and $s_{i_1} s_{i_2} \ldots s_{i_l}$ is a reduced word of w.

This classification can be identified with the Tits distance obtained from a Bruhat decomposition of GL_n. In particular, being in w relative position does not depend on the choice of reduced word of w. If $\mathscr{F} \overset{w}{\sim} \mathscr{F}'$, then, for each choice of reduced word $(i_1, \ldots, i_l)$ for w, there exists a unique sequence of flags $(\mathscr{G}_k)_{k=0}^l$ that relate the two flags $\mathscr{F}$ and $\mathscr{F}'$.

We can now translate the local quiver representation data associated with a Type 2 column into relative position relations between flags. Suppose the pair of flags to the left of a Type 2 column is $(\mathscr{L}_0, \mathscr{L}^0)$, and the pair of flags to the right of a Type 2 column is $(\mathscr{R}_0, \mathscr{R}^0)$. If there is a crossing in the bottom region at the ith gap, counting from the bottom in both the front and the GP graph, then, from the exactness condition at the crossing, we obtain the constraints

$$(4\text{-}4) \qquad \mathscr{L}_0 \overset{s_i}{\sim} \mathscr{R}_0 \quad \text{and} \quad \mathscr{L}^0 = \mathscr{R}^0.$$

Similarly, if there is a crossing in the top region at the ith gap, counting from the top in the front projection or counting from the bottom in the GP graph, then the exactness condition at the crossing yields

$$(4\text{-}5) \qquad \mathscr{L}_0 = \mathscr{R}_0 \quad \text{and} \quad \mathscr{L}^0 \overset{s_{n-i}}{\sim} \mathscr{R}^0.$$

Since crossings in Type 2 columns correspond to vertical edges in the GP graph, we can infer the relative position relation between pairs of flags from the GP graph as well.

For a Type 3 column, the pairs of flags on the (Type 1 column on the) left and on the (Type 1 column on the) right are not in the same ambient vector space, as the dimensions of the two vector spaces differ by one. Instead, there is a linear map *from* the ambient vector space for the pair of flags on the left *to* the ambient vector space for the pair of flags on the right. This linear map is injective if the lollipop is white and it is surjective if the lollipop is black. Let us investigate how the two pairs of flags are related.

Suppose first that the lollipop is white, so that the linear map $h \colon \mathbb{C}^{n-1} \to \mathbb{C}^n$ between the two (middle) adjacent ambient vector spaces is injective. Given any flag $\mathscr{F} = (0 \subset \mathscr{F}_1 \subset \cdots \subset \mathscr{F}_{n-1} = \mathbb{C}^{n-1})$ in $\mathbb{C}^{n-1}$, we can use h to naturally extend it to a flag $h(\mathscr{F})$ in $\mathbb{C}^n$ by defining

$$h(\mathscr{F}) := (0 \subset h(\mathscr{F}_1) \subset h(\mathscr{F}_2) \subset \cdots \subset h(\mathscr{F}_{n-1}) \subset \mathbb{C}^n).$$

This extension from $(\mathscr{L}_0, \mathscr{L}^0)$ to $(h(\mathscr{L}_0), h(\mathscr{L}^0))$ can be achieved geometrically by a sequence of RII moves that pulls the left cusp upward in the front projection. Indeed, consider the local example in Figure 38.

The green maps in the bottom region define the extension $h(\mathscr{L}_0)$. By the exactness of the quadrilaterals in the top region, the red maps define the extension $h(\mathscr{L}^0)$. In particular, the extensions $h(\mathscr{L}_0)$ and $h(\mathscr{L}^0)$ are completely determined by the original data of the quiver representations.

Now, with these extensions defined, it follows that, if there is a white lollipop emerging in the ith gap with $0 \leq i \leq n-1$, counting from below in the GP graph,[11] then the corresponding relative position conditions are

$$(4\text{-}6) \qquad h(\mathscr{L}_0) \overset{s_{n-1}\cdots s_{i+1}}{\sim} \mathscr{R}_0 \quad \text{and} \quad h(\mathscr{L}^0) \overset{s_{n-1}\cdots s_{n-i}}{\sim} \mathscr{R}^0.$$

---

[11]The case $i = 0$ is a lollipop at the bottom, and $i = n-1$ is a lollipop at the top.
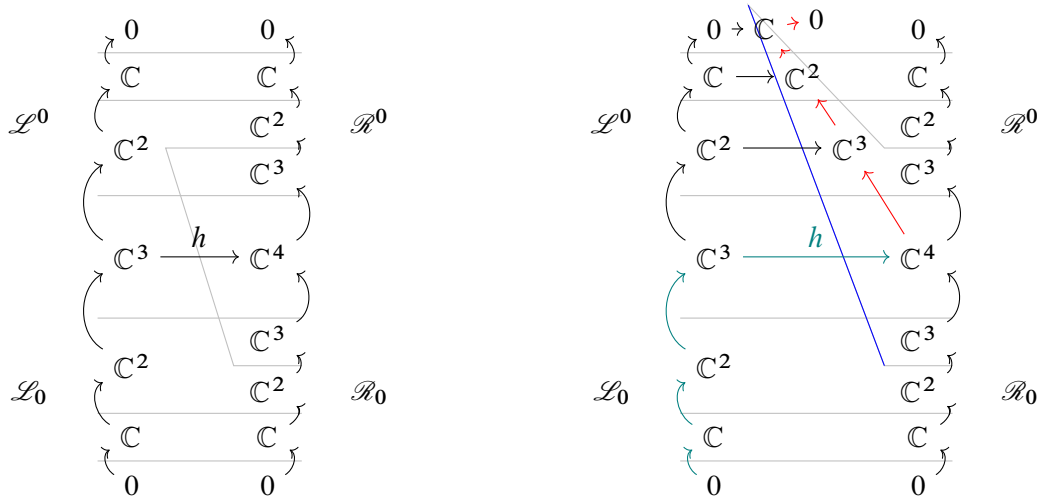
Figure 38: Pulling up a left cusp.

Suppose that there is a black lollipop, and thus the linear map between the two ambient vector spaces $h\colon \mathbb{C}^n \to \mathbb{C}^{n-1}$ is surjective. Then, given any flag $\mathscr{F} = (0 \subset \mathscr{F}_1 \subset \cdots \subset \mathscr{F}_{n-1} = \mathbb{C}^{n-1})$ in $\mathbb{C}^{n-1}$, we consider $h^{-1}(\mathscr{F}_i)$ and insert $\ker(h)$ in front of it so as to form a flag in $\mathbb{C}^n$:

$$h^{-1}(\mathscr{F}) := (0 \subset \ker(h) \subset h^{-1}(\mathscr{F}_1) \subset \cdots \subset h^{-1}(\mathscr{F}_{n-1}) = \mathbb{C}^n).$$

Similar to the white lollipop case, the extension of $(\mathscr{R}_0, \mathscr{R}^0)$ to $(h^{-1}(\mathscr{R}_0), h^{-1}(\mathscr{R}^0))$ can be achieved geometrically by a sequence of RII moves that pulls the right cusp downward in the front projection, as depicted in Figure 39. Note that the green maps in bottom region define the extension $h^{-1}(\mathscr{R}_0)$, whereas the red maps in the top region define the extension $h^{-1}(\mathscr{R}^0)$.
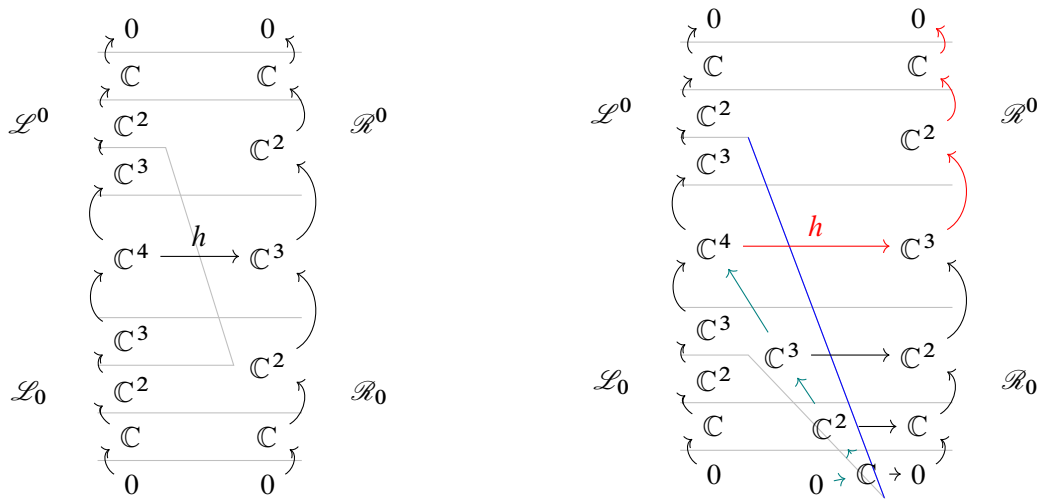
Figure 39: Pulling down a right cusp.

It follows that, if there is a black lollipop occurring in the ith gap with $0 \leq i \leq n$, counting from below in the GP graph, then the corresponding relative position conditions are

$$(4\text{-}7) \qquad \mathscr{L}_0 \overset{s_{i-1}\cdots s_1}{\sim} h^{-1}(\mathscr{R}_0) \quad \text{and} \quad \mathscr{L}^0 \overset{s_{n-i+1}\cdots s_1}{\sim} h^{-1}(\mathscr{R}^0).$$

In summary, given a GP graph $\mathbb{G}$, we can divide $\mathbb{G}$ into columns of three types such that every consecutive pair of non-Type 1 columns is separated by a Type 1 column and every consecutive pair of Type 1 columns is separated by a non-Type 1 column. In between Type 3 columns there is a unique ambient vector space $V_i = \mathbb{C}^n$ for some n, and they are linked by linear maps $h_i \colon V_{i-1} \to V_i$ that are either injective with a 1–dimensional cokernel or surjective with a 1–dimensional kernel. The above discussion proves:

Lemma 4.2 *For a GP graph $\mathbb{G}$ with a decomposition into columns as above, the moduli space $\mathcal{M}_1(\Lambda)$ can be described by the following data:*

(1) *a pair of flags in V_i for each Type 1 column contained in the V_i part of $\mathbb{G}$;*

(2) *for each Type 2 column, the neighboring flags satisfy the relative position condition according to (4-4) and (4-5);*

(3) *for each Type 3 column, the neighboring flags satisfy the relative position condition according to (4-6) and (4-7);*

where we quotient this data by the equivalence relation $(\mathscr{F}, h) \sim (\mathscr{F}', h')$ for a collection of elements $g_i \in \mathrm{GL}(V_i)$ such that $h_i \circ g_{i-1} = g_i \circ h'_i$.

In the flag description of Lemma 4.2, the rank-1 local system Φ on Λ can be constructed by taking quotients of consecutive vector subspaces in each flag and then gluing them along strands of Λ at crossings and cusps in the same manner as before. Note that, in this context, only (4-1) is used when gluing these rank-1 local systems at crossings because all linear maps near a crossing are now inclusions of vector subspaces. In particular, the surjective maps in the top region of the front projection are now turned into inclusions of kernels.

4.1.2 Description for (-1)–closures Finally, there is another description of $\mathcal{M}_1(\Lambda)$ and $\mathfrak{M}(\Lambda, \mathrm{T})$ as moduli space of configurations of flags, which aligns better when comparing with the flag moduli of the weaves $\mathfrak{w}(\mathbb{G})$. In that latter case, there will be only one ambient vector space. The description in Lemma 4.2, which is associated to the specific front $\mathfrak{f}(\mathbb{G})$, after using RII and RIII moves, can be shown to be equivalent to a description with a unique ambient (top-dimensional) vector space. Indeed, rather than using flags from different ambient spaces with varying dimensions, we can perform additional RII and RIII moves to push strands like the blue one in Figure 38 all the way to the left and push strands like the blue one in Figure 39 all the way to the right (see also Lemma 3.22). This will extend all flags from all Type 1 columns to flags in $\mathbb{C}^h$, where h is the total number of horizontal lines in the GP graph $\mathbb{G}$. Moreover, these flags will satisfy the relative position conditions imposed by the external weave lines of the initial weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ or, equivalently, the cyclic positive braid word $\beta = \beta(\mathbb{G})$ for which Λ is its (-1)–closure. In this context, [Shende et al. 2017, Proposition 1.5], or Proposition 4.1, reads:

Lemma 4.3 Let $\beta = (i_1, i_2, \ldots, i_l) \in \mathrm{Br}_h^+$ be a positive braid word on h strands and Λ be the Legendrian link associated to the front given by the (-1)–closure of β. Then

$$\mathcal{M}_1(\Lambda) \cong \left\{ (\mathscr{F}_0, \mathscr{F}_1, \mathscr{F}_2, \ldots, \mathscr{F}_l) \,\middle|\, \begin{array}{l} \mathscr{F}_i \text{ is a flag in } \mathbb{C}^h \text{ for all } i, \\ \mathscr{F}_0 \overset{s_{i_1}}{\sim} \mathscr{F}_1 \overset{s_{i_2}}{\sim} \cdots \overset{s_{i_l}}{\sim} \mathscr{F}_l = \mathscr{F}_0 \end{array} \right\} \Big/ \mathrm{PGL}_h \,.$$

4.2 Factoriality property

In the upcoming construction of a cluster $\mathcal{A}$–structure for the moduli space $\mathfrak{M}(\Lambda, T)$, we shall need that the coordinate ring $\mathcal{O}(\mathfrak{M}(\Lambda, T))$ is a unique factorization domain (aka. factorial). This can be a subtle condition to verify and thus we provide in this section an argument that the condition of Δ–completeness of the braid $\beta(\mathbb{G})$, as introduced in Section 2.5, is sufficient for factoriality. Note that all shuffle graphs have $\beta(\mathbb{G})$ be a Δ–complete braid, and thus the rings $\mathcal{O}(\mathfrak{M}(\Lambda(\mathbb{G}), T))$ are factorial if $\mathbb{G}$ is shuffle.

Proposition 4.4 Let $\mathbb{G}$ be a GP graph with $\beta(\mathbb{G})$ a Δ–complete braid. Then the moduli space $\mathfrak{M}(\Lambda(\mathbb{G}), T)$ is an affine variety whose coordinate ring is factorial.

Proof Since the moduli space $\mathfrak{M}(\Lambda, T)$ is a Legendrian invariant, without loss of generality we can turn Λ into the (-1)–closure of an n–stranded positive braid $\beta(\mathbb{G}) = \Delta\gamma$ and use the description from Section 4.1.2. Let us first consider the case where the set T of marked points can be arranged into a configuration with one marked point per level along a vertical line between Δ and γ. (It follows that $|T| = n$.) This case is depicted as follows:
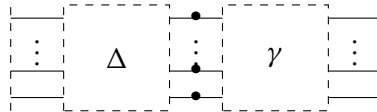
Let B_+ and B_- be the Borel subgroups of PGL_n of upper-triangular and lower-triangular matrices, respectively. We can exhaust the PGL_n–action on flag configurations by fixing the two flags at the two ends of Δ to be the two unique flags stabilized by B_+ and B_-, respectively, while requiring that the decoration on the flag $\mathscr{F}_l$ at the dashed line (after γ on the right or before Δ on the left) be the standard one, ie mapping $\bar{e}_i$ to 1 for each consecutive quotient $\mathrm{Span}\{e_1, \ldots, e_i\}/\mathrm{Span}\{e_1, \ldots, e_{i-1}\} \cong \mathrm{Span}(\bar{e}_i)$.

Let $(i_1, \ldots, i_l)$ be a positive word for the positive braid γ such that $\beta(\mathbb{G}) = \Delta\gamma$. Let us record a flag as a matrix with row vectors such that the span of the last k row vectors give the k–dimensional subspace in the flag. Then $\mathscr{F}_l$ can be recorded by the permutation matrix w_0. Starting from the flag $\mathscr{F}_l$, the flags $\mathscr{F}_{l-1}, \mathscr{F}_{l-2}, \ldots$ to the left of $\mathscr{F}_l$ can then be given by

$$\mathscr{F}_k = B_{i_{k+1}}(z_{k+1}) B_{i_{k+2}}(z_{k+2}) \cdots B_{i_l}(z_l) w_0.$$

In the end, we need $\mathscr{F}_0$ to be the standard flag

$$0 \subset \mathrm{Span}\{e_n\} \subset \mathrm{Span}\{e_{n-1}, e_n\} \subset \cdots \subset \mathrm{Span}\{e_2, \ldots, e_n\} \subset \mathbb{C}^n,$$

which is equivalent to requiring that $B_{i_1}(z_1)B_{i_2}(z_2)\cdots B_{i_k}(z_l)w_0$ be upper-triangular. This shows that $\mathfrak{M}(\Lambda(\mathbb{G}), T)$ is isomorphic to the braid variety $X(\beta(\mathbb{G}), w_0)$ from [Casals et al. 2020], which has a factorial coordinate ring.[12]

Now let us consider the case of an arbitrary number of marked points. Let us start with the set T having one marked point per level, as in the case above. Suppose m of the marked points share the same link component; then we can move these marked points along that link component until they get inside an horizontal interval with no crossings or cusps. Then these marked points are just changing decorations on the same underlying 1–dimensional quotient of consecutive vector spaces of the same flag. Thus, we can extract a $(\mathbb{C}^\times)^{m-1}$–torus factor and replace these marked points with one marked point. By doing this for each link component, we can reduce T to a set T' with one marked point per link component, and conclude that

$$\mathfrak{M}(\Lambda, T) \cong \mathfrak{M}(\Lambda, T') \times (\mathbb{C}^\times)^{n-N}$$

as affine varieties, where N is the number of link components in $\Lambda = \Lambda(\mathbb{G})$. This implies that

$$\mathbb{O}(\mathfrak{M}(\Lambda, T)) \cong \mathbb{O}(\mathfrak{M}(\Lambda, T')) \otimes \mathbb{C}[t_i^{\pm 1}]_{i=1}^{n-N}.$$

If there is an element in $\mathbb{O}(\mathfrak{M}(\Lambda, T'))$ admitting two nonequivalent factorizations, then these two factorizations are still valid and nonequivalent in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$, contradicting the fact that $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ is factorial. Thus, we can conclude that $\mathbb{O}(\mathfrak{M}(\Lambda, T'))$ is factorial when T' consists of one marked point per link component. In general, for any set T'' with at least one marked point per link component, we can implement the same argument above and write

$$\mathfrak{M}(\Lambda, T'') \cong \mathfrak{M}(\Lambda, T') \times (\mathbb{C}^\times)^{|T''|-N}.$$

Algebraically, this implies that

$$\mathbb{O}(\mathfrak{M}(\Lambda, T'')) \cong \mathbb{O}(\mathfrak{M}(\Lambda, T')) \otimes \mathbb{C}[t_i^{\pm 1}]_{i=1}^{|T''|-N}.$$

Again, since $\mathbb{O}(\mathfrak{M}(\Lambda, T'))$ is factorial, so is the tensor product $\mathbb{O}(\mathfrak{M}(\Lambda, T')) \otimes \mathbb{C}[t_i]_{i=1}^{|T''|-N}$. Given that $\mathbb{O}(\mathfrak{M}(\Lambda, T')) \otimes \mathbb{C}[t_i^{\pm 1}]_{i=1}^{|T''|-N}$ is a localization of this factorial tensor product, it is factorial as well. $\square$

4.3 Moduli spaces for the Lagrangian $L(\mathfrak{w}(\mathbb{G}))$

The moduli spaces $\mathcal{M}_1(\Lambda)$ and $\mathfrak{M}(\Lambda, T)$ depend only on the Legendrian isotopy type of Λ. In particular, if $\Lambda = \Lambda(\mathbb{G})$ is a GP link, then these moduli spaces are invariant under square moves and other combinatorial equivalences of the GP graph $\mathbb{G}$ which preserve the Legendrian isotopy class of Λ. The GP graph also provides the information of an embedded exact Lagrangian filling for $\Lambda(\mathbb{G})$, namely the exact Lagrangian filling $L = L(\mathbb{G})$ described by the initial weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$. The Guillermou–Jin–Treumann map of [Jin and Treumann 2017], or [Ekholm et al. 2016; Casals and Ng 2022], imply that there are open embeddings

$$H^1(L; \mathbb{C}^\times) \to \mathcal{M}_1(\Lambda), \quad H^1(L, T; \mathbb{C}^\times) \to \mathfrak{M}(\Lambda, T),$$

---

[12]We thank Eugene Gorsky for an explanation of why this is the case. See also upcoming work of the first author with Gorsky and coauthors, where this is written in detail.

whose domains parametrize (decorated) $\mathbb{C}$–local systems on L (with decoration T), and the map is essentially the microlocalization functor. These open torus charts $(\mathbb{C}^{\times})^{b_1(L)}$ and $(\mathbb{C}^{\times})^{b_1(L,T)}$ can be described in terms of flags if the Lagrangian filling L is obtained from a weave, as explained in [Casals and Zaslow 2022]; we shall use it in the proof of Theorem 1.1. The definition of $\mathcal{M}_1(\mathfrak{w})$ from [Casals and Zaslow 2022] is as follows:

Definition 4.5 Let $\mathfrak{w} \subset \mathbb{R}^2$ be a weave. By definition, the *total flag moduli space* $\widetilde{\mathcal{M}}_1(\mathfrak{w})$ associated to $\mathfrak{w}$ comprises tuples of flags, as follows:

(i) There is a flag $\mathscr{F}^{\bullet}(F)$ assigned to each face F of the weave $\mathfrak{w}$, ie to each connected component of $\mathbb{R}^2 \setminus \mathfrak{w}$.

(ii) For each pair of adjacent faces $F_1, F_2 \subset \mathbb{R}^2 \setminus \mathfrak{w}$, sharing an s_i–edge, their two associated flags $\mathscr{F}^{\bullet}(F_1)$ and $\mathscr{F}^{\bullet}(F_2)$ are in relative position $s_i \in S_n$, ie they must satisfy

$$\mathscr{F}_j(F_1) = \mathscr{F}_j(F_2) \quad \text{for } 0 \leq j \leq N \text{ with } j \neq i \qquad \text{and} \qquad \mathscr{F}_i(F_1) \neq \mathscr{F}_i(F_2).$$

The group PGL_n acts on the space $\widetilde{\mathcal{M}}_1(\mathfrak{w})$ simultaneously. By definition, the *flag moduli space* of the weave $\mathfrak{w}$ is the quotient stack $\mathcal{M}_1(\mathfrak{w}) := \widetilde{\mathcal{M}}(\mathfrak{w})/\mathrm{PGL}_n$.

By Section 4.1.1, $\mathcal{M}_1(\mathfrak{w})$ is an open subspace of $\mathcal{M}_1(\Lambda)$ via restriction to the boundary. Indeed, since the weaves $\mathfrak{w}$ are free weaves [Casals and Zaslow 2022, Section 7.1.2], the data of flags at the boundary of the initial weave uniquely determines the flags at each face of $\mathfrak{w}$. (This fact can also be verified combinatorially.) It follows from [Casals and Zaslow 2022] that $\mathcal{M}_1(\mathfrak{w})$ are complex tori $\mathcal{M}_1(\mathfrak{w}) \cong (\mathbb{C}^{\times})^{\dim \mathcal{M}_1(\Lambda)}$, and thus these moduli spaces of flags associated to the initial weave $\mathfrak{w}$ are natural candidates for an initial cluster chart in the moduli space $\mathcal{M}_1(\Lambda)$ for a GP link Λ. (These complex tori are indeed the images of the Guillermou–Jin–Treumann maps.) The definition of candidate cluster $\mathscr{X}$–variables will be the subject of the next subsection.

The decorated version of the flag moduli $\mathcal{M}_1(\mathfrak{w})$, which we denote by $\mathfrak{M}(\mathfrak{w}, T)$, is naturally defined by adding a framing away from T along the boundary $\partial L(\mathfrak{w}) = \Lambda$. It also follows that $\mathfrak{M}(\mathfrak{w}, T)$ is naturally an open torus chart in $\mathfrak{M}(\Lambda, T)$. The corresponding definition of the candidate cluster $\mathscr{A}$–variables is undertaken in Section 4.6.

4.4 Microlocal monodromies: unsigned candidate $\mathscr{X}$–variables

Let us consider the open toric chart $\mathcal{M}_1(\mathfrak{w}) \subset \mathcal{M}_1(\Lambda)$ from Section 4.3. We now build a function

$$X_{\gamma} \colon \mathcal{M}_1(\mathfrak{w}) \to \mathbb{C}$$

associated to each Y–cycle γ, generalizing [Casals and Zaslow 2022, Section 7] — see also [Shende et al. 2017, Section 5.1] — to our context. First we observe that the data of $\mathcal{M}_1(\mathfrak{w})$ associates a flag $\mathscr{F}$ in each connected component of the complement of $\mathfrak{w}$ in $\mathbb{R}^2$. We associate the 1–dimensional vector space
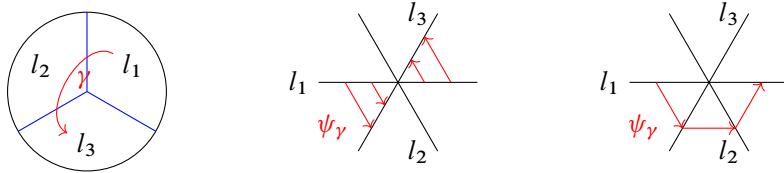
Figure 40: A weave for the unique filling of the max tb unknot and microlocal parallel transports from its sheaf quantization.

$\mathscr{F}_i/\mathscr{F}_{i-1}$ to the ith sheet in the lift of each connected component. Then, across the lifts of each weave line, we define two linear isomorphisms

(4-8)
$$\psi_+ : \frac{\mathscr{L}_i}{\mathscr{L}_{i-1}} \hookrightarrow \frac{\mathscr{L}_{i+1}}{\mathscr{L}_{i-1}} = \frac{\mathscr{R}_{i+1}}{\mathscr{R}_{i-1}} \twoheadrightarrow \frac{\mathscr{R}_{i+1}}{\mathscr{R}_i},$$

$$\psi_- : \frac{\mathscr{R}_i}{\mathscr{R}_{i-1}} \hookrightarrow \frac{\mathscr{R}_{i+1}}{\mathscr{R}_{i-1}} = \frac{\mathscr{L}_{i+1}}{\mathscr{L}_{i-1}} \twoheadrightarrow \frac{\mathscr{L}_{i+1}}{\mathscr{L}_i}.$$

Note that $\psi_\pm$ are isomorphisms because $\mathscr{L}$ and $\mathscr{R}$ are in s_i–transverse position, as they are separated by a weave line labeled with s_i. Now, given a loop γ on L, we may perturb it so that it intersects with any lifts of weave lines transversely. Then, by composing several of the isomorphisms $\psi_\pm$ above and their inverses, we obtain a linear automorphism for each generic fiber along γ. Since each generic fiber is a 1–dimensional vector space, we can represent this linear automorphism by a nonzero scalar ψ_γ. This nonzero scalar ψ_γ is also known as the *microlocal monodromy* of the sheaf moduli space $\mathcal{M}_1(\mathfrak{w})$ along γ. However, the microlocal monodromies ψ_γ do not naturally give rise to a local system on L,[13] as the following illustrates:

Example 4.6 Consider the weave with a unique trivalent vertex, which depicts a Lagrangian 2–disk filling, as drawn in blue in Figure 40, left. According to the definition of $\mathcal{M}_1(\mathfrak{w})$, there is a flag $l_i \subset \mathbb{C}^2$ in each of the three sectors, and they are pairwise transverse. Let γ be a curve on $L(\mathfrak{w}) \cong \Lambda(\mathfrak{w})$ which, under the front projection, goes from the lower sheet to the upper sheet and then back to the lower sheet; see again Figure 40, left. By definition, the microlocal parallel transport ψ_γ should be the map in Figure 40, center, which is the linear map that projects parallel to the line l_2. Consider the lift ξ of a loop that goes around a trivalent weave vertex in $\mathbb{R}^2$, which is a double cover for the projection onto the weave plane. Without loss of generality, let us suppose ξ starts at the lower sheet in the sector containing l_1. The parallel transport along ξ is then the composition of the three linear projections, as in Figure 40, right, which is equal to the linear map $v \mapsto -v$ on l_1. In other words, $\psi_\xi = -1$. However, ξ is a contractible cycle on $L(\mathfrak{w})$ and thus the microlocal monodromy assignment $\xi \mapsto \psi_\xi$ cannot be a local system on $L(\mathfrak{w})$.
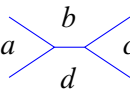
Let us now specialize to our situation, with $\mathbb{G}$ a GP graph and $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ its initial weave. By Section 3, there is a distinguished linearly independent subset $\mathfrak{S}(\mathbb{G}) \subset H_1(L(\mathfrak{w}(\mathbb{G})))$ of $\mathbb{L}$–compressible cycles

---

[13]They give a *twisted* local system as in [Guillermou 2023, Part 13], or a twisted flat connection as in [Gaiotto et al. 2013, Part 10].

parametrized by the sugar-free hulls in the GP graph. For each element in $\mathfrak{S}(\mathbb{G})$, we choose a Y–tree representative γ, which exists by Lemma 3.29, and define

$$X_\gamma := -\psi_\gamma.$$

These functions shall become our cluster $\mathscr{X}$–variables, once signs are fixed and Theorem 1.1 is proven. Note that, since we can isotope the Y–tree γ to a short l–cycle, ie an equivalent monochromatic edge, we may use it to compute X_γ explicitly, as follows. In a neighborhood of a short l–cycle, labeled with the permutation s_i, a point in the flag moduli $\mathcal{M}_1(\mathfrak{w})$ is specified by the data of a quadruple of flags. Each of these flags has the same subspaces $\mathscr{F}^j$ in each region for $j \neq i$, and for $j = i$ we additionally require the data in each region of a line l in the 2–dimensional space $V := \mathscr{F}^{i+1}/\mathscr{F}^{i-1}$. This is the data of four lines $a, b, c, d \subset V$. The function X_γ is then equal to the cross-ratio

$$X_\gamma = \langle a, b, c, d \rangle = -\frac{a \wedge b}{b \wedge c} \cdot \frac{c \wedge d}{d \wedge a}.$$

The definition of X_γ, following [Fock and Goncharov 2006b; Shende et al. 2019; Casals and Zaslow 2022], is not particularly new. It is also possible to define X_γ directly and combinatorially from the Y–trees, in line with [Casals and Zaslow 2022, Section 7]. The fact that these functions $\{X_\gamma\}$ transform according to an $\mathscr{X}$–mutation formula under a square-face mutation is due to [Shende et al. 2019], and under the more general weave mutation due to [Casals and Zaslow 2022]. Indeed, let $\Gamma = \{\gamma_i\}$ be a maximal collection of Y–trees in $\mathfrak{w}(\mathbb{G})$ which are linearly independent in $H_1(L(\mathfrak{w}(\mathbb{G})))$, $Q(\Gamma)$ be their (algebraic) intersection quiver, and $X_\Gamma = \{X_{\gamma_i}\}$ be a labeling of each vertex of the quiver. Then it is shown in [Casals and Zaslow 2022, Section 7.2.2] that weave mutation at one such Y–tree $\gamma \in \Gamma$ induces a quiver mutation of $Q(\Gamma)$ at the vertex associated to γ, and the set of variables X_Γ changes according to a cluster $\mathscr{X}$–mutation.

Defining these candidate cluster $\mathscr{X}$–variables is relatively useless for the purpose of proving existence of cluster structures: the variables X_γ do *not* extend to global in $\mathcal{M}_1(\Lambda)$ in general and we cannot deduce the existence of a cluster $\mathscr{X}$–structure merely from constructing this initial seed $(Q(\Gamma), X_\Gamma)$. Moreover, in general there could be many choices of Γ for a fixed general weave $\mathfrak{w}$, and it is not known whether different choices yield equivalent, or even quasiequivalent, cluster seeds. It thus becomes crucial to construct cluster $\mathscr{A}$–variables for $\mathfrak{M}(\Lambda, T)$, ideally in a symplectic invariant manner, as we will do in a moment. By [Berenstein et al. 2005], a cluster $\mathscr{A}$–structure can be shown to exist once the necessary properties of the candidate $\mathscr{A}$–variables are proven. As a byproduct, Corollary 1.2 then deduces the existence of the cluster $\mathscr{X}$–structure on $\mathcal{M}_1(\Lambda)$ where the variables are microlocal monodromies.

4.5 Collections of sign curves: fixing signs

Let $\mathbb{G}$ be a GP graph, $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ its initial weave and $L := L(\mathfrak{w})$ its initial filling, and T a set of marked points in $\Lambda(\mathbb{G}) = \partial L$. Let us denote the set of lifts of trivalent weave vertices on L by $P \subset L$. It follows from Sections 4.3 and 4.4 that each point of the flag moduli $\mathcal{M}_1(\mathfrak{w})$ defines a rank 1 local system

on $L \setminus P$ with -1 monodromy around each point in P. In this subsection, we describe a way to add signs to monodromies to obtain a (noncanonical) isomorphism between $\mathcal{M}_1(\mathfrak{w})$ and $\mathrm{Loc}_1(L)$. This is a combinatorial expression of the fact that, in our case, global sections of the Kashiwara–Schapira stack are (canonically) isomorphic to the category of twisted local systems and (noncanonically) also isomorphic to the category of local systems. In terms of weave combinatorics, we proceed as follows:

Definition 4.7 A *sign curve* is an unoriented curve on the weave surface L that intersects the lifts of weave lines transversely and whose endpoints lie in the set $P \sqcup T$. By definition, a collection C of sign curves on L is *coherent* if each point in P is incident to one and only one sign curve in C, and all curves in C intersect transversely.

We record sign curves on L by drawing dotted curves on $\mathbb{R}^2$ in juxtaposition with the weave $\mathfrak{w}$ and labeling the indices of the sheets they are on.

Fix a coherent set C of sign curves on L. For any path γ on L, we may perturb γ so that it intersects elements of C transversely. Then we redefine the parallel transport along γ to be the microlocal parallel transport ψ_γ multiplied by a factor of -1 whenever the curve γ passes through a sign curve in C. Since each branch point of L is incident to one and only one sign curve, this new parallel transport corrects the monodromy around each point in P to be 1, defining an isomorphism

$$\Phi_C : \mathcal{M}_1(\mathfrak{w}) \to \mathrm{Loc}_1(L) \cong H^1(L; \mathbb{C}^\times) \cong (\mathbb{C}^\times)^{b_1(L)}.$$

In fact, we can do better than an arbitrary isomorphism $\mathcal{M}_1(\mathfrak{w}) \xrightarrow{\cong} \mathrm{Loc}_1(L)$. From Section 4.4, our candidates for cluster $\mathscr{X}$–variables are of the form $-\psi_\gamma$ for initial absolute cycles $\gamma \in \mathfrak{S}(\mathbb{G})$, and we can in fact incorporate this extra sign in front of ψ_γ into the set of coherent sign curves.

Definition 4.8 A coherent set C of sign curves on L is said to be *compatible* if, for all initial absolute cycles $\gamma \in \mathfrak{S}(\mathbb{G})$,

$$\Phi_C(p)(\gamma) = X_\gamma(p) := -\psi_\gamma(p) \quad \text{for all } p \in \mathcal{M}_1(\mathfrak{w}).$$

For the initial free weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ constructed from a GP graph $\mathbb{G}$, we can find a compatible set of sign curves as follows. First, we observe that all trivalent weave vertices of $\mathfrak{w}$ occur near the boundary of the weave. Thus, at each trivalent weave vertex, two of the three adjacent sectors are facing away from the weave: we will draw our sign curves inside these two sectors. Next, we break the weave $\mathfrak{w}$ down into weave columns, and, by Section 3, trivalent weave vertices only occur inside Type 2 columns.

Let us further classify Type 2 columns into two types: a Type 2 column is said to be *critical* if it is the rightmost Type 2 column that contains part of an initial cycle $\gamma \in \mathfrak{S}(\mathbb{G})$; it is said to be *noncritical* otherwise. By construction, each critical Type 2 column has a unique initial cycle γ that ends there.

If a Type 2 column is noncritical, we draw a sign curve in either sector on either sheet, and then lead it towards the boundary of L; once it gets within a collar neighborhood of the boundary $\partial L = \Lambda$, the sign

curve will follow along Λ until it reaches a marked point. (Such a marked point exists because we have at least one marked point per link component.)

If a Type 2 column is critical, we consider the unique initial cycle γ that ends at this Type 2 column. We compute the product of all the signs γ has picked up along all the previous trivalent weave vertices. If the product is 1, we add a sign curve c on the appropriate sheet of either of the two sectors so that γ intersects with c nontrivially. If the product is -1, we add a sign curve c on the appropriate sheet of either of the two sectors so that γ intersects with c trivially. By doing so, we guarantee that Φ_C maps γ to $X_\gamma = -\psi_\gamma$, as desired.

Thus, a compatible set C of sign curves exists in our setting, and we can explicitly identify $\mathcal{M}_1(\mathfrak{w})$ with the moduli space $\mathrm{Loc}_1(L)$ of rank 1 local systems on L. This identification allows us to interpret the cluster $\mathcal{X}$–variables X_γ as actual monodromies of local systems along the initial cycles γ, and also fixing the necessary signs for the upcoming constructions.

Proposition 4.9 *If $\mathfrak{w}$ admits a compatible set of sign curves and $\mathfrak{w}'$ is mutation equivalent to $\mathfrak{w}$, then $\mathfrak{w}'$ also admits a (noncanonical) compatible set of sign curves.*

Proof   Both weave equivalences and weave mutations are local operations on the weave. Therefore, it suffices to verify that compatible sets of sign curves can be constructed locally, before and after such local operations. That is, locally in a neighborhood where the weave equivalence or mutation is going to be performed, we want to argue that any given compatible set of sign curves on that piece of the initial weave — before an equivalence or mutation — we can construct a compatible set of sign curves afterwards, locally on that piece of the weave after the operation.

For weave equivalences, Moves I, IV and V in Definition 3.2 do not involve any trivalent weave vertices. Thus, sign curves that pass through any of these local pictures can be carried through these equivalences using planar homotopies. In contrast, Move II (the push-through move), III and VI do involve trivalent weave vertices. In the case of Move VI, the weave lines lift to sheets that are not adjacent to each other; therefore, the weave line with no trivalent vertex (yellow in Figure 12) can be ignored when studying the set of compatible sign curves, reducing to the constant case of a trivalent vertex. By Remark 3.3, Move III is a concatenation of Moves I and II. Thus, the only weave equivalence move that remains to be studied is Move II, which will be discussed in a moment. For weave mutations, it suffices to check mutations along short I–cycles, since any Y–tree is weave equivalent to a short I–cycle by Proposition 3.5. In conclusion, we need to study compatible sets of sign curves locally near a push-through and a weave mutation.

A priori, we must study sign curves that arrive at the trivalent vertices from different faces of the weave; a face being any connected component of the complement of the weave lines. That said, if a sign curve is incident to a trivalent weave vertex, we can apply a planar homotopy to the sign curve so that it arrives at the trivalent vertex from any of the another faces near the trivalent vertex. This is depicted in Figure 41. Therefore, it suffices to study the case that a sign curve arrives at a trivalent only from one of the three
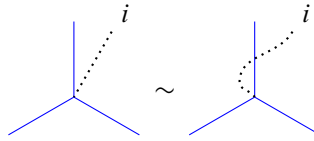
Figure 41: Applying a planar homotopy to a sign curve so that it arrives at the trivalent vertex from a different face. (The index i can be either 1 or 2.)

faces near the trivalent vertex. (If the curve arrived from another face, we could homotope the curve and, locally in a small neighborhood around the vertex, have it arrive from another face.)

With this reduction, it suffices to study compatible sets of sign curves locally near a push-through and a weave mutation which arrive from one face (of our choice) at each trivalent vertex. Up to symmetry, there are only three cases to check, shown in Figure 42. The figure illustrates how to resolve the problem at hand: for each set of compatible sign curves before the equivalence or mutation (on the left of each diagram), we can construct a set of compatible sign curves afterwards (on the right of each diagram). □

4.6 Microlocal merodromies: candidate cluster $\mathscr{A}$–variables

This subsection addresses the construction of what shall become the cluster $\mathscr{A}$–variables on the moduli space $\mathfrak{M}(\Lambda, T)$ for a GP link $\Lambda = \Lambda(\mathbb{G})$. In the previous subsection, we explained that cluster $\mathscr{X}$–variables were indexed by certain *absolute* cycles $\gamma \in H_1(L)$ in the Lagrangian filling $L = L(\mathfrak{w}(\mathbb{G}))$ and X_γ was a natural rational function with a symplectic origin: the microlocal monodromy along γ of the sheaf associated with the weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$.

Now, the new idea is that cluster $\mathscr{A}$–variables $\{A_\eta\}$ will be indexed by certain *relative* cycles $\eta \in H_1(L \setminus T, \Lambda \setminus T)$ and the functions $A_\eta \colon \mathfrak{M}(\Lambda, T) \to \mathbb{C}$ will be defined by what we call the *microlocal merodromy* along η. Intuitively, this merodromy along η is constructed as a microlocal parallel transport along η. Here are the details.

4.6.1 Microlocal merodromy Let $\mathbb{G}$ be an GP graph, $\Lambda = \Lambda(\mathbb{G})$ be its GP link, and T be a collection of marked points on Λ with at least one marked point per link component. Fix a compatible set C of sign curves and let $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ be the initial weave. The flag moduli $\mathfrak{M}(\mathfrak{w}, T)$ is an open subset of the moduli space $\mathfrak{M}(\Lambda, T)$, and every point in this open subset defines a local system, via the identification $\Phi_C \colon \mathcal{M}_1(\mathfrak{w}) \xrightarrow{\cong} \mathrm{Loc}_1(L)$ in Section 4.5, together with a framing (trivialization) of the rank-1 local system Φ on the connected components of $\Lambda \setminus T = (\partial L) \setminus T$.
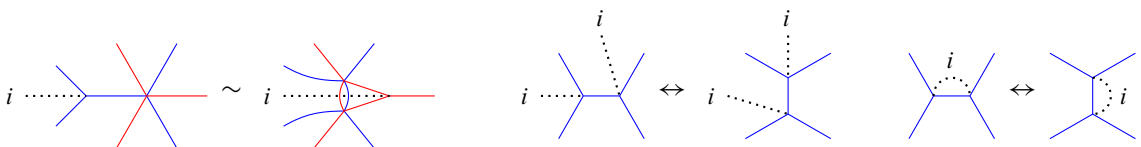
Figure 42: Existence of compatible sets of sign curves before and after weave equivalences (left) and weave mutations (center and right). (The index i can be either 1 or 2.)
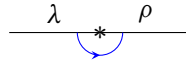
Figure 43: A marked point $t \in T$, with decorations λ and ρ to the left and right, and the boundary of a half-disk neighborhood U_t.

The framing data defines a special vector $\phi_x \in \Phi_x$ at any point $x \in \Lambda \setminus T$. Given an oriented curve $\eta \subset L$ with both the source point $s = \partial_- \eta$ and the target point $t = \partial_+ \eta$ contained inside $\Lambda \setminus T$, we can parallel transport ϕ_s from the source s to the target t along η, obtaining a nonzero vector in $\eta(\phi_s) \in \Phi_t$. The ratio $\eta(\phi_s)/\phi_t$ is a nonzero number A_η, which defines a $\mathbb{C}^\times$–valued function on $\mathfrak{M}(\mathfrak{w}, T)$. This can be naturally generalized to relative 1–cycles $\eta \in H_1(L \setminus T, \Lambda \setminus T)$.

Definition 4.10 The function $A_\eta \colon \mathfrak{M}(\mathfrak{w}, T) \to \mathbb{C}^\times$ is said to be the *microlocal merodromy* along the oriented curve η.

Since $\mathfrak{M}(\mathfrak{w}, T)$ is an open subset of $\mathfrak{M}(\Lambda, T)$, A_η can also be viewed as a rational function on $\mathfrak{M}(\Lambda, T)$. Note that, a priori, A_η might not extend to a regular function on $\mathfrak{M}(\Lambda, T)$. We emphasize that the decorations in $\mathfrak{M}(\Lambda, T)$ are needed in order to define A_η, and thus microlocal merodromies cannot be defined in $\mathcal{M}_1(\Lambda)$.

The microlocal merodromies associated to relative cycles coming from marked points are nonvanishing. Indeed, for each marked point $t \in T$, we pick a small half-disk neighborhood U_t of t, as in Figure 43, and define $\xi_t := \partial U_t$. By definition,

$$A_t := A_{\xi_t} = \frac{\xi_t(\lambda)}{\rho} \neq 0.$$

In particular, by using the ratio $\xi_t(\lambda)/\rho$, we can extend A_t to a global invertible function on the entire moduli space $\mathfrak{M}(\Lambda, T)$. (This property does not in general hold for A_η if η is an arbitrary relative cycle in $H_1(L \setminus T, \Lambda \setminus T)$.) Now consider the exact sequence of lattices

$$0 \to \mathbb{Z} \xrightarrow{i} \bigoplus_{t \in T} \mathbb{Z}\xi_t \to H_1(L \setminus T, \Lambda \setminus T) \xrightarrow{\pi} H_1(L, \Lambda) \to 0,$$

where $i(1) := \sum_{t \in T} \xi_t$. This exact sequence implies the following two corollaries:

Corollary 4.11 $\prod_{t \in T} A_t = 1$ and hence A_t is a unit in $\mathcal{O}(\mathfrak{M}(\Lambda, T))$ for every $t \in T$.

Proof This follows from the fact that $\sum_{t \in T} \xi_t = 0$ in $H_1(L \setminus T, \Lambda \setminus T)$. $\qquad \square$

Corollary 4.12 If $\eta_1, \eta_2 \in H_1(\Sigma \setminus T, \Lambda \setminus T)$ satisfy $\pi(\eta_1) = \pi(\eta_2)$, then A_{η_1} and A_{η_2} are related to each other by a Laurent monomial in the variables A_t for $t \in T$.

Proof This follows from the fact that $\ker \pi = \mathrm{Span}\{\xi_t \mid t \in T\}$. $\qquad \square$

The latter corollary starts to hint at the quasicluster equivalence that appears if different basis completions in $H_1(\Sigma \setminus T, \Lambda \setminus T)$ are chosen, as the former corollary indeed hints at the fact that A_t for $t \in T$ are a type of frozen variables.

4.6.2 Crossing values The next aim is to compute A_η for curves whose support is transverse to a weave $\mathfrak{w}$. Consider the commutative diagram of vector space inclusions

$$
\begin{array}{ccc}
 & V_n & \\
\nearrow & & \nwarrow \\
V_w & & V_e \\
\nwarrow & & \nearrow \\
 & V_s &
\end{array}
$$

and assume that $0 \to V_s \to V_w \oplus V_e \to V_n \to 0$ is exact. Let $\alpha_s, \alpha_w, \alpha_e$ and α_n be nonzero top-dimensional (volume) forms in V_s, V_w, V_e and V_n, respectively. Then we write $\alpha_w = \beta_w \wedge \alpha_s$ and $\alpha_e = \beta_e \wedge \alpha_s$ for some forms β_w and β_e.

Definition 4.13 In the context of a diagram as above, we define

$$
\alpha_w \overset{\alpha_s}{\wedge} \alpha_e := \beta_w \wedge \beta_e \wedge \alpha_s.
$$

The top form $\alpha_w \overset{\alpha_s}{\wedge} \alpha_e$ is nonzero on V_n and does not depend on the choice of β_w or β_e. By definition, the ratio $(\alpha_w \overset{\alpha_s}{\wedge} \alpha_e)/\alpha_n$ is said to be the *crossing value* of the quadruple of top forms $\alpha_s, \alpha_w, \alpha_e, \alpha_n$.

Let us describe how to use crossing values to compute merodromies along planar relative cycles. Consider a flag $\mathscr{F} = (0 \subset \mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}_n = \mathbb{C}^n)$ with a choice of $\phi_i \neq 0 \in \mathscr{F}_i/\mathscr{F}_{i-1}$ for all $i \in [1, n]$. The choice of such $\phi = (\phi_i)$ for $i \in [1, n]$ is said to be a *framing* for the flag $\mathscr{F}$. Given such *framed flag* $(\mathscr{F}, \phi)$, we can construct top forms $\alpha_i \in \bigwedge^i \mathscr{F}_i$ for $i \in [1, n]$, by first lifting each ϕ_j to a vector in $\widetilde{\phi}_j \in \mathscr{F}_j$ for $j \in [1, n]$, and then taking ordered wedges, leading to the forms

$$
(4\text{-}9) \qquad\qquad \alpha_i := \widetilde{\phi}_i \wedge \widetilde{\phi}_{i-1} \wedge \cdots \wedge \widetilde{\phi}_1 \quad \text{for } i \in [1, n].
$$

Note that each form α_i is independent of the choice of lifts.

Definition 4.14 Given a flag $\mathscr{F}$, a collection $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ of nonvanishing forms $\alpha_i \in \bigwedge^i \mathscr{F}_i$ for $i \in [1, n]$ is said to be a *decoration* on the flag $\mathscr{F}$. A flag with a decoration is referred to as a *decorated flag*.

Note that we can reverse the construction above and recover a framing from a decoration. Thus, framings $(\phi_1, \ldots, \phi_n)$ and decorations $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ of a flag $\mathscr{F}$ are equivalent pieces of data. By definition, two decorated (or framed) flags $(\mathscr{L}, \alpha)$ and $(\mathscr{R}, \beta)$ are in s_i–transverse position if the underlying flags $\mathscr{L} \overset{s_i}{\sim} \mathscr{R}$ are in s_i–transverse position, ie s_i–transversality does not see decorations (or framings).

Suppose $(\mathscr{L}, \lambda)$ and $(\mathscr{R}, \rho)$ are two framed flags such that $\mathscr{L} \overset{s_i}{\sim} \mathscr{R}$. Let α and β be the decorations constructed from λ and ρ, respectively. Consider the parallel transport maps $\psi_\pm$ defined in (4-8). The images $\psi_+(\lambda_i)$ and $\psi_-(\rho_i)$ are readily computed in terms of decorations as follows:

Lemma 4.15 *We have*

$$
\psi_+(\lambda_i) = \frac{\alpha_i \overset{\alpha_{i-1}}{\wedge} \beta_i}{\beta_{i+1}} \rho_{i+1} \quad \text{and} \quad \psi_-(\rho_i) = \frac{\beta_i \overset{\beta_{i-1}}{\wedge} \alpha_i}{\alpha_{i+1}} \lambda_{i+1}.
$$

Proof Let $\widetilde{\lambda}_i$ and $\widetilde{\rho}_{i+1}$ be lifts of λ_i and ρ_{i+1}. By construction, the framing $\psi_+(\lambda_i)$ can obtained as follows. First, lift $\lambda_i \in \mathcal{L}_i / \mathcal{L}_{i-1}$ to a vector $\widetilde{\lambda}_i \in \mathcal{L}_i$ and consider this vector as $\widetilde{\lambda}_i \in \mathcal{L}_{i+1}$ via the inclusion $\mathcal{L}_i \subset \mathcal{L}_{i+1}$. Then, using $\mathcal{L}_{i+1} = \mathcal{R}_{i+1}$, we can view $\widetilde{\lambda}_i \in \mathcal{R}_{i+1}$ and thus finally $\psi_+(\lambda_i) = \pi(\widetilde{\lambda}_i)$, where $\pi : \mathcal{R}_{i+1} \to \mathcal{R}_{i+1}/\mathcal{R}_i$ is the quotient map. Each of $\psi_+(\lambda_i)$ and ρ_{i+1} is a (volume) 1–form on $\mathcal{R}_{i+1}/\mathcal{R}_i$, and can be pulled back via π to 1–forms in $\mathcal{R}_{i+1}$. By wedging these forms with (any) top form in $\mathcal{R}_i$, such as β_i, we obtain the top forms $\widetilde{\lambda}_i \wedge \beta_i$ and $\widetilde{\rho}_{i+1} \wedge \beta_i$. Since we wedged both $\psi_+(\lambda_i)$ and ρ_{i+1} with the same form β_i, their ratios are equal:

$$\frac{\psi_+(\lambda_i)}{\rho_{i+1}} = \frac{\widetilde{\lambda}_i \wedge \beta_i}{\widetilde{\rho}_{i+1} \wedge \beta_i}.$$

By (4-9), we also have $\alpha_i = \widetilde{\lambda}_i \wedge \alpha_{i-1}$ and $\beta_{i+1} = \widetilde{\rho}_{i+1} \wedge \beta_i$. Therefore,

$$\frac{\psi_+(\lambda_i)}{\rho_{i+1}} = \frac{\widetilde{\lambda}_i \wedge \beta_i}{\widetilde{\rho}_{i+1} \wedge \beta_i} = \frac{\alpha_i \overset{\alpha_{i-1}}{\wedge} \beta_i}{\beta_{i+1}}.$$

The equality for ψ_- is obtained similarly. $\qquad\square$

Remark 4.16 By Lemma 4.15, the inverses of $\psi_\pm$ are computed analogously. Namely,

$$\psi_+^{-1}(\rho_{i+1}) = \frac{\beta_{i+1}}{\alpha_i \overset{\alpha_{i-1}}{\wedge} \beta_i} \lambda_i \quad \text{and} \quad \psi_-^{-1}(\lambda_{i+1}) = \frac{\alpha_{i+1}}{\beta_i \overset{\beta_{i-1}}{\wedge} \alpha_i} \rho_i.$$

Now suppose $\eta \subset \Sigma(\mathfrak{w})$ is a lift of a planar curve in $\mathbb{R}^2$ to the weave front. Then it defines a partial cross-section of the weave surface, where η passes through a collection of (framed) flags $\mathcal{L} = \mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_l = \mathcal{R}$. For each flag $\mathcal{F}_i$ for $0 < i < l$ we choose a sequence of top forms $\alpha_j = (\alpha_{i,j})$. Since the parallel transport along η consists of compositions of linear isomorphisms like the maps $\psi_\pm$ in Lemma 4.15, or their inverses, Lemma 4.15 allows us to compute A_η.

Example 4.17 Consider the cross-section of a weave surface depicted in Figure 44, and let η be the blue relative cycle. The sequences of top forms α and δ are determined by the decorations λ and ρ, respectively. The tuples of top forms β and γ are chosen arbitrarily. Note that $\alpha_0, \beta_0, \gamma_0$ and δ_0 are trivial. Let us denote the $\psi_\pm$ maps associated to each of the three crossings by ${}_i\psi_\pm$, where $i \in [1, 3]$, $i = 1$ being associated to the leftmost crossing, $i = 2$ to the center crossing and $i = 3$ to the rightmost crossing.

By definition, the microlocal merodromy along η is

$$A_\eta = \frac{\psi_\eta(\lambda_2)}{\rho_3} = \frac{({}_3\psi_+ \circ {}_2\psi_+ \circ {}_1\psi_-^{-1})(\lambda_2)}{\rho_3}.$$

By Lemma 4.15, each of the microlocal merodromies ${}_i\psi_\pm$ and their inverses are computed as

$$A_\eta = \frac{\psi_\eta(\lambda_2)}{\rho_3} = \frac{\gamma_2 \overset{\gamma_1}{\wedge} \delta_2}{\delta_3} \cdot \frac{\beta_1 \wedge \gamma_1}{\gamma_2} \cdot \frac{\alpha_2}{\beta_1 \wedge \alpha_1} = \frac{\beta_1 \wedge \delta_2}{\delta_3} \cdot \frac{\alpha_2}{\beta_1 \wedge \alpha_1}.$$
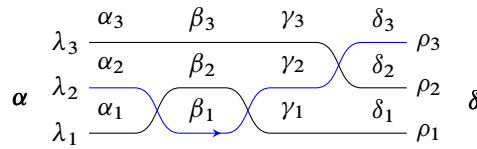
Figure 44: Computation of a merodromy.

Two observations based on this computation: First, the right-hand side of this expression shows that A_η depends on the underlying undecorated flag associated with the β, but it is invariant under any nonzero rescaling of the decoration β. This is a general fact. Namely, the function A_η does depend on the intermediate flags between the two flags at the endpoints; nevertheless, it does not depend on the decorations of these intermediate flags.

Second, a reason for the decoration γ not appearing in the computation of A_η above is that the flag associated to γ is uniquely determined by the flags associated to β and δ. Observe that in the case that the slice along η yields a reduced braid word, the intermediate flags are uniquely determined by the flags at the endpoints and thus the microlocal merodromy only depends on the decorated flags at the endpoints. For more general computations, the study of microlocal merodromies involves understanding properties, such as regularity, of products of crossing values and inverses thereof.

In general, a microlocal merodromy A_η will be expressed in terms of ratios of crossing values, and is only a *rational* function. Nevertheless, certain choices of η within the initial weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ yield *regular* functions. Indeed, let us consider the following special family of merodromies. By Section 3.6, each face f of $\mathbb{G}$ has an associated naive relative cycle η_f in $H_1(L \setminus T, \Lambda \setminus T)$. Let $A_f := A_{\eta_f}$ be the microlocal merodromy of this naive relative cycle. Since $\{\partial f\}$ is a basis of $H_1(L)$, it follows from Poincaré duality that $\{\pi(\eta_f)\}$ is a basis of $H_1(L, \Lambda)$, where $\pi : H_1(L \setminus T, \Lambda \setminus T) \to H_1(L, \Lambda)$ is the natural projection map. By Corollary 4.12, we conclude that different choices of η_f only change A_f by a multiple of units.

Proposition 4.18 *Let f be a face in a GP graph $\mathbb{G}$. Then the microlocal merodromy*

$$A_f : \mathfrak{M}(\Lambda(\mathbb{G}), T) \to \mathbb{C}$$

is a regular function.

Proof Suppose that the face $f \in \mathbb{G}$ corresponds to the kth gap in $\mathbb{G}$, counting from the bottom. Then the associated relative cycle η_f can be written as

$$\eta_f = \sum_{i=1}^{k} \xi_i,$$

where the relative cycles ξ_i are described as follows. Consider the braid word $\mathfrak{w}_{0,n}$; then ξ_i is the relative cycle given by the ith strand in $\mathfrak{w}_{0,n}$, counting from the bottom on the left, when considered as a slice of

the weave $\mathfrak{w}(\mathbb{G})$ along η_f. The following figure depicts such ξ_i for $n = 4$:
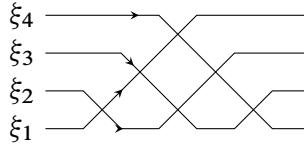
Now, the microlocal parallel transport along ξ_i, from the ith strand on the left to the $(n-i+1)^{\text{st}}$ strand on the right, can be computed via Lemma 4.15. In particular, each microlocal merodromy A_{ξ_i} has contributions from $i - 1$ crossings because $\mathfrak{w}_{0,n}$ is a half-twist, and thus it is obtained after composing $i - 1$ instances of $\psi_{\pm}^{-1}$. In such a slice spelling $\mathfrak{w}_{0,n}$, let $(\mathscr{L}, \alpha)$ be the decorated flag at the left endpoints and $(\mathscr{R}, \beta)$ the decorated flag at the left endpoints. In line with Example 4.17, we obtain

$$A_{\xi_i} = \frac{\gamma \wedge \beta_{n-i}}{\beta_{n-i+1}} \cdot \frac{\alpha_i}{\gamma \wedge \alpha_{i-1}} \quad \text{for } i \in [1, k],$$

where γ is a nonzero vector in the line $\mathscr{L}_i \cap \mathscr{R}_{n-i+1}$. Note that the formula for A_{ξ_i} does not actually depend on γ. In fact, since $\mathfrak{w}_{0,n}$ is a reduced word, we have $\alpha_{i-1} \wedge \beta_{n-i+1} \neq 0$. Thus, after wedging both $\gamma \wedge \beta_{n-i}$ and β_{n-i+1} with α_{i-1}, the expression for A_{ξ_i} above reads

$$\begin{aligned}
A_{\xi_i} &= \frac{\alpha_{i-1} \wedge \gamma \wedge \beta_{n-i}}{\alpha_{i-1} \wedge \beta_{n-i+1}} \cdot \frac{\alpha_i}{\gamma \wedge \alpha_{i-1}} = (-1)^{i-1} \frac{\alpha_{i-1} \wedge \gamma \wedge \beta_{n-i}}{\alpha_{i-1} \wedge \beta_{n-i+1}} \cdot \frac{\alpha_i}{\alpha_{i-1} \wedge \gamma} \\
&= (-1)^{i-1} \frac{\alpha_{i-1} \wedge \gamma \wedge \beta_{n-i}}{\alpha_{i-1} \wedge \beta_{n-i+1}} \cdot \vartheta = (-1)^{i-1} \frac{\vartheta \cdot (\alpha_{i-1} \wedge \gamma) \wedge \beta_{n-i}}{\alpha_{i-1} \wedge \beta_{n-i+1}} \\
&= (-1)^{i-1} \frac{\alpha_i \wedge \beta_{n-i}}{\alpha_{i-1} \wedge \beta_{n-i+1}} \quad \text{for } i \in [1, k],
\end{aligned}$$

where we have denoted by $\vartheta \in \mathbb{C}^{\times}$ the unique nonzero scalar such that $\alpha_i = \vartheta \cdot (\alpha_{i-1} \wedge \gamma)$. In conclusion, we obtain

$$A_f = \prod_{i=1}^{k} A_{\xi_i} = (-1)^{k(k-1)/2} \cdot \frac{\alpha_k \wedge \beta_{n-k}}{\beta_n}.$$

By definition, $\beta_n \neq 0$ is nonzero and therefore A_f is a regular function for each face f. $\qquad\square$

Remark 4.19 Microlocal merodromies can also be used to define the frozen cluster $\mathscr{X}$–variables associated to the relative cycles in $H_1(L, T)$ that are not in the image of $H_1(L)$. In the moduli space $\mathcal{M}_1(\Lambda, T)$, the microlocal merodromy allows one to compare framings at the endpoints of the relative cycles, which are marked points T where the (stalk of the microlocal) local system has been trivialized.

4.7 Vanishing of microlocal merodromies and flag relative positions

In this subsection we study the vanishing loci of the microlocal merodromies $A_f : \mathfrak{M}(\Lambda(\mathbb{G}), T) \to \mathbb{C}$ associated to faces $f \subset \mathbb{G}$ of a GP graph $\mathbb{G}$. The key technical result, Proposition 4.23, relates the vanishing loci of microlocal merodromies associated to different faces of the graph $\mathbb{G}$. This result is

crucial to deduce the necessary properties of these candidate cluster $\mathcal{A}$–variables, such as regularity, and conclude Theorem 1.1.

Thus far, we have parametrized the relative position of a pair of flags in $\mathbb{C}^n$ by the symmetric group S_n. This relative position is invariant under the diagonal GL_n action, and hence is also in bijection with GL_n–orbits in $\mathcal{B}(n) \times \mathcal{B}(n)$. The inclusion relation on closures of these orbits defines a partial order, called the *Bruhat order*, on S_n, ie $u \leq v$ if $\mathbb{O}_u \subset \bar{\mathbb{O}}_v$. Combinatorially, the Bruhat order can be computed through set comparison.

Definition 4.20 For two equal-sized subsets $I = \{i_1 < i_2 < \cdots < i_m\}$ and $J = \{j_1 < j_2 < \cdots < j_m\}$ of $\{1, \ldots, n\}$, we define $I \leq J$ if $i_k \leq j_k$ for all $1 \leq k \leq m$. By definition, for two permutations u and v of S_n, $u \leq v$ in the *Bruhat order* if and only if $\{u(1), \ldots, u(m)\} \leq \{v(1), \ldots, v(m)\}$ for all $1 \leq m < n$.

By Section 4.1, the moduli $\mathfrak{M}(\Lambda(\mathbb{G}), T)$ can be understood in terms of tuples of flags, with maps between them and incidence constraints. The flags can be read directly from the front $\mathfrak{G}$. In particular, for a Type 1 column of $\mathbb{G}$, there exists a unique pair of (decorated) flags $\mathscr{F}_0$ and $\mathscr{F}^0$; see Figure 37. In the points of the open torus chart $\mathfrak{M}(\mathfrak{w}, T) \subset \mathfrak{M}(\Lambda(\mathbb{G}), T)$, these two flags $\mathscr{F}_0$ and $\mathscr{F}^0$ are in w_0–relative position, but in general the relative position between $\mathscr{F}_0$ and $\mathscr{F}^0$ at another point of $\mathfrak{M}(\Lambda(\mathbb{G}), T)$ might vary. The dependence of $\mathscr{F}_0$ and $\mathscr{F}^0$ on the point $p \in \mathfrak{M}(\Lambda(\mathbb{G}), T)$ will be denoted by $\mathscr{F}_0(p)$ and $\mathscr{F}^0(p)$.

By Proposition 4.18, the microlocal merodromy A_i associated to the ith gap of a Type 1 column, counting from below in the GP graph, is a regular function on $\mathfrak{M}(\Lambda, T)$. Moreover, it can be expressed as $(\alpha_i \wedge \beta_{n-i})/\beta_n$, up to a multiple of units, where α and β are decorations on the pair of flags $\mathscr{F}_0$ and $\mathscr{F}^0$ placed at the bottom and the top of that Type 1 column, respectively. In particular, the restriction of $A_i|_{\mathfrak{M}(\mathfrak{w}, T)}$ to the open torus chart $\mathfrak{M}(\mathfrak{w}, T) \subset \mathfrak{M}(\Lambda(\mathbb{G}), T)$ is a nonvanishing function. The following lemma shows that we can describe the vanishing locus of this microlocal merodromy $A_i : \mathfrak{M}(\Lambda(\mathbb{G}), T) \to \mathbb{C}$ in terms of the relative position between the two flags $\mathscr{F}_0$ and $\mathscr{F}^0$:

Lemma 4.21 *Let $\mathbb{G}$ be a GP graph, $(\mathscr{F}_0, \alpha), (\mathscr{F}^0, \beta)$ the pair of decorated flags associated with a Type 1 column C and A_i the ith microlocal merodromy associated to C for $i \in [1, n]$. Consider a point $p \in \mathfrak{M}(\Lambda(\mathbb{G}), T)$ and the permutation $w \in S_n$ such that $\mathscr{F}_0(p) \overset{w}{\sim} \mathscr{F}^0(p)$. Then $A_i(p) = 0$ if and only if $w \leq s_i w_0$ in the Bruhat order.*

Proof Without loss of generality, we may assume that the decorations α and β are proportional to

$$(e_{w(1)}, e_{w(1)} \wedge e_{w(2)}, \ldots, e_{w(1)} \wedge e_{w(2)} \wedge \cdots \wedge e_{w(n)}) \quad \text{and} \quad (e_1, e_1 \wedge e_2, \ldots, e_1 \wedge e_2 \wedge \cdots \wedge e_n),$$

respectively. From this we know that $A_i = 0$ if and only if

$$e_{w(1)} \wedge e_{w(2)} \wedge \cdots \wedge e_{w(i)} \wedge e_1 \wedge e_2 \wedge \cdots \wedge e_{n-i} = 0,$$

which is equivalent to saying that

$$\{w(1), w(2), \ldots, w(i)\} \cap \{1, 2, \ldots, n-i\} \neq \varnothing.$$

Now note that, for the permutation $v = s_i w_0$, all $\{v(1), \ldots, v(m)\}$ are maximal sets with respect to the linear order on $\{1, \ldots, n\}$ except when $m = i$, where

$$\{v(1), \ldots, v(i)\} = \{n, n-1, \ldots, n-i+2, n-i\}.$$

If $w \leq v$, then $\{w(1), w(2), \ldots, w(i)\} \leq \{n, n-1, \ldots, n-i+2, n-i\}$. This implies that, among $w(1), w(2), \ldots, w(i)$, some index no greater than $n-i$ must have appeared. Therefore, we have $\{w(1), w(2), \ldots, w(i)\} \cap \{1, 2, \ldots, n-i\} \neq \varnothing$ and hence $A_i = 0$.

Conversely, if $w \not\leq v$, then we must have

$$\{w(1), w(2), \ldots, w(i)\} = \{n, n-1, \ldots, n-i+1\},$$

which implies that $\{w(1), w(2), \ldots, w(i)\} \cap \{1, 2, \ldots, n-i\} = \varnothing$ and hence $A_i \neq 0$. □

Lemma 4.21 shows that, in order to study whether the microlocal merodromies A_f vanish or not, it suffices to consider the relative position between the pair of flags in a Type 1 column that contains part of the face f. We use the following simple lemma in the proof of Proposition 4.23, through Lemma 4.24:

Lemma 4.22 *Let* $u, v \in S_n$ *and consider three flags* $\mathscr{F}$, $\mathscr{F}'$ *and* $\mathscr{F}''$ *such that* $\mathscr{F} \overset{u}{\sim} \mathscr{F}' \overset{v}{\sim} \mathscr{F}''$. *Let* l *denote the length function on* S_n *and, for any* $w \in S_n$, *we denote by* $\underline{w}$ *the positive braid represented by a* (*equivalently any*) *reduced word of* w. *Then the following holds*:

(1) *If* $l(uv) = l(u) + l(v)$, *then* $\mathscr{F} \overset{uv}{\sim} \mathscr{F}''$.

(2) *In general, if* $\mathscr{F} \overset{w}{\sim} \mathscr{F}'$, *then* $w \leq \mathrm{Dem}(\underline{u}\,\underline{v})$ *in the Bruhat order.*[14]

Proof (1) follows from the standard fact that, for Bruhat cells, if $l(uv) = l(u) + l(v)$ then $(BuB)(BvB) = BuvB$. For (2), at the level of Bruhat cells we know that $(Bs_iB)(Bs_iB) = Bs_iB \sqcup B$; this is equivalent to saying that, if $\mathscr{F} \overset{s_i}{\sim} \mathscr{F}' \overset{s_i}{\sim} \mathscr{F}''$, then there are two possible relative position relations between $\mathscr{F}$ and $\mathscr{F}''$: either $\mathscr{F} \overset{s_i}{\sim} \mathscr{F}''$ or $\mathscr{F} = \mathscr{F}''$. For both cases, the relative position is at most $\mathrm{Dem}(\underline{s_i}\,\underline{s_i}) = s_i$. The general statement follows from the well-definedness of the Demazure product. □

For notational convenience, for $1 \leq i \leq j < n$, let us define the permutation

$$
w_{[i,j]} := \begin{matrix}
 & & & s_{n-1} & & & \\
 & & s_{n-2} & & s_{n-2} & & \\
 & & \iddots & & \ddots & & \\
 & s_2 & s_2 & s_2 & \cdots & s_2 & \\
 s_1 & s_1 & s_1 & & & & s_1
\end{matrix}
\qquad \text{and} \qquad \bar{w} := w_0 w^{-1} w_0.
$$

$$\underbrace{}_{\text{delete all } s_1\text{'s from the } i^{\text{th}} \text{ to } j^{\text{th}} \text{ copy}}$$

---

[14]Dem denotes the Demazure product on positive braids.

In this notation, $s_i w_0 = w_{[i,i]} = \bar{w}_{[i,i]}$. In terms of set comparison, $u \leq w_{[i,j]}$ in the Bruhat order if and only if, for all $n - j \leq l \leq n - i$,

(4-10)
$$\{u(1), u(2), \ldots, u(l)\} \leq \{i \leq \cdots\},$$

where $\{i \leq \cdots\}$ means the set of the appropriate size (say of size k) consisting of the greatest $k - 1$ elements in $\{1, \ldots, n\}$ together with the element i.

Finally, the core of this subsection is the following result, which states that we can use lollipop reactions (Definition 2.9) to keep track of the relative position conditions on flags and, in turn, understand vanishing conditions for microlocal merodromies associated to faces:

Proposition 4.23 *Let $\mathbb{G}$ be a GP graph and $f, g \subset \mathbb{G}$ two faces. Suppose that g is selected in a lollipop reaction initiated from a lollipop in f. Then $A_f = 0$ implies $A_g = 0$.*

As discussed in Section 2.6, the scanning wall in a lollipop reaction moves to the right if the lollipop is white and to the left if the lollipop is black. By Lemma 4.21, at the starting point, $A_f = 0$ implies that the flags at the two ends of the wall are at most $w_{[i,i]} = \bar{w}_{[i,i]}$ apart, where i is the index of the gap between the two adjacent horizontal lines, counting from below in the GP graph. This is schematically illustrated as

The heart of the argument is proving that, in the case of a white (resp. black) lollipop reaction, as the wall scans to the right (resp. left), the flags at the two ends of the wall will be at most $w_{[i,j]}$ (resp. $\bar{w}_{[i,j]}$) apart, where $[i, j]$ is the interval containing the indices of the gaps that the wall crosses (counting from below in the GP graph). Due to symmetry, we will only prove Proposition 4.23 for white lollipop reactions; the proof for the case of black lollipop reactions is completely symmetric. Let us start with the following lemma:

Lemma 4.24 *Let $\mathbb{G}$ be a GP graph and consider a Type 3 column with a black lollipop (we shift the indices on the right to match the Coxeter generators), as follows:*
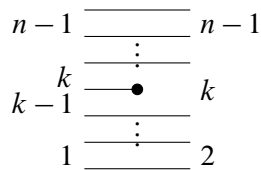
Let $(\mathcal{L}_0, \mathcal{L}^0)$ be the pair of flags to the left and $(\mathcal{R}_0, \mathcal{R}^0)$ be the pair of flags to the right. Suppose that $\mathcal{L}_0 \overset{u}{\sim} \mathcal{L}^0$ and $\mathcal{R}_0 \overset{v}{\sim} \mathcal{R}^0$. (Hence, $h^{-1}(\mathcal{R}_0) \overset{v}{\sim} h^{-1}(\mathcal{R}^0)$.) Then, in the Bruhat order,

$$u \geq s_{k-1}s_{k-2}\ldots s_1 v s_1 s_2 \ldots s_{n-k}.$$

Proof By construction, $h^{-1}(\mathscr{R}_0)$ and $h^{-1}(\mathscr{R}^0)$ share the same 1–dimensional subspace. Therefore, $v(1) = 1$, which implies that $vs_1s_2 \ldots s_{n-k}$ is reduced. Lemma 4.22(1) implies that

$$\mathscr{L}_0 \overset{s_{k-1} \ldots s_1}{\sim} h^{-1}(\mathscr{R}_0) \overset{vs_1 \ldots s_{n-k}}{\sim} \mathscr{L}^0,$$

where the first relative position is given by the Type 3 column requirement.

Let us record a permutation $w \in S_n$ as an n–tuple $(w(1), w(2), \ldots, w(n))$. Since $v(1) = 1$, we may assume that $v = (1, v(2), v(3), \ldots, v(n))$. Then

$$vs_1s_2 \ldots s_{n-k} = (v(2), v(3), \ldots, v(n-k+1), 1, v(n-k+2), \ldots, v(n)).$$

Note that left multiplication by s_i interchanges the entries i and $i+1$. From (the proof of) Lemma 4.22(2), we know that, when multiplying s_i on the left of a permutation w, there is only one possible relative position $s_i w$ if i is on the left of $i+1$, and there can be two possible relative positions w and $s_i w$ if i is on the right of $i+1$, in which case $s_i w < w$. Thus, performing all the left multiplications $s_1, \ldots, s_{k-1}$ in turn on $vs_1 \ldots s_{n-k}$ yields the smallest relative position relation, and hence $u \geq s_{k-1} \ldots s_1 vs_1 \ldots s_{n-k}$, as claimed. $\qquad\square$

Proof of Proposition 4.23 It suffices to argue in the case of a white lollipop reaction, by symmetry. We inductively verify the claim that, as the wall scans from left to right, the relative position between the pair of flags remains at most $w_{[i,j]}$.

Suppose that the wall scanning is passing through a column of Type 2 or 3. Let $(\mathscr{L}_0, \mathscr{L}^0)$ be the pair of flags on the left of this column and $(\mathscr{R}_0, \mathscr{R}^0)$ be the pair of flags on the right of this column. Suppose that the wall on the left goes across the interval $[i, j]$. Then, by assumption, $\mathscr{L}_0 \overset{u}{\sim} \mathscr{L}^0$ for $u \leq w_{[i,j]}$. Let v be the permutation such that $\mathscr{R}_0 \overset{v}{\sim} \mathscr{R}^0$.

Let us start with the hardest case, namely a Type 3 column with a black lollipop at the kth gap with $i < k \leq j$, as depicted in Figure 45, left. By a shift of indices, we can view v as an element in $S_{[2,n]}$, the permutation group that acts on the set $[2, n] = \{2, 3, \ldots, n\}$. We want to prove that $v \leq w_{[i,j-1]} \in S_{[2,n]}$. By shifting the indices of the Coxeter generators in (4-10), $v \leq w_{[i,j-1]} \in S_{[2,n]}$ if and only if, for all $n - j \leq l \leq n - i - 1$,

$$\{v(2), v(3), \ldots, v(l+1)\} \leq \{i + 1 \leq \cdots\}.$$

Let us proceed by contradiction: Suppose $v \nleq w_{[i,j-1]} \in S_{[2,n]}$; then there must exist some l with $n - j \leq l \leq n - i - 1$ such that

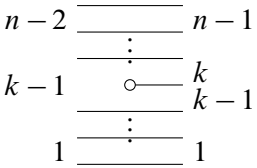$$\{v(2), v(3), \ldots, v(l+1)\} = \{a, a + \cdots, \ldots\},$$

where $a > i + 1$ is the smallest element in the set on the right. To deduce a contradiction, it suffices to prove that $\tilde{v} := s_{k-1} \ldots s_1 vs_1 \ldots s_{n-k} \nleq w_{[i,j]} \in S_{[1,n]}$, since Lemma 4.24 states that $\tilde{v}$ is the smallest possible relative position between $\mathscr{L}_0$ and $\mathscr{L}^0$. Direct computation yields that

$$\tilde{v} = (v(2)', v(3)', \ldots, v(n-k+1)', k, v(n-k+2)', \ldots, v(n)'),$$
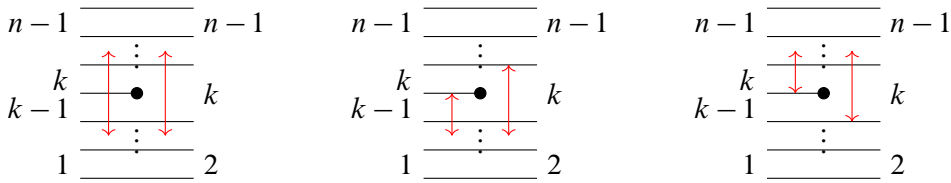
Figure 45: Three of the cases in the proof of Proposition 4.23.

where

$$m' := \begin{cases} m-1 & \text{if } m \le k, \\ m & \text{if } m > k. \end{cases}$$

If $l \le n - k$, then

$$\{\tilde{v}(1), \tilde{v}(2), \dots, \tilde{v}(l)\} = \{v(2)', v(3)', \dots, v(l+1)'\} = \{a', a' + \cdots, \dots\} \not\le \{i \le \cdots\},$$

where the last $\not\le$ relation is because $a' \ge a - 1 > i$. This shows that $\tilde{v} \not\le w_{[i,j]}$.

Otherwise, if $l > n - k$, then

$$\{\tilde{v}(1), \tilde{v}(2), \dots, \tilde{v}(l+1)\} = \{v(2)', v(3)', \dots, v(n-k+1)', k, v(n-k+2)', v(l+1)'\}$$
$$= \{k, a', a' + \cdots, \dots\} \not\le \{i \le \cdots\},$$

where the last $\not\le$ relation is because both a' and k are greater than i.

There are three more special cases to consider for Type 3 columns with black lollipops, two of them depicted in Figure 45, center and right, as well as the case where neither $k - 1$ nor k is contained in $[i, j]$.

For the case of Figure 45, center, we want to prove that, if $u \le w_{[i,k-1]} \in S_{[1,n]}$, then $v \le w_{[i,k-1]} \in S_{[2,n]}$. We proceed by contradiction again. Suppose $v \not\le w_{[i,k-1]}$; then there must exist some l with $n - k \le l \le n - i - 1$ such that

$$\{v(2), v(3), \dots, v(l+1)\} = \{a, a + \cdots, \dots\},$$

where $a > i + 1$ is the smallest element on the right. By the same argument, we see that

$$\{\tilde{v}(1), \tilde{v}(2), \dots, \tilde{v}(l+1)\} = \{v(2)', v(3)', \dots, v(n-k+1)', k, v(n-k+2)', v(l+1)'\}$$
$$= \{k, a', a' + \cdots, \dots\} \not\le \{i \le \cdots\}.$$

This shows that $\tilde{v} \not\le w_{[i,k-1]} \in S_{[1,n]}$.

For the case of Figure 45, right, we want to prove that, if $u \le w_{[k,j]} \in S_{[1,n]}$, then $v \le w_{[k-1,j-1]} \in S_{[2,n]}$; a proof by contradiction works again, as follows. Suppose $v \not\le w_{[k-1,j-1]}$; then there must exist some l with $n - j \le l \le n - k$ such that

$$\{v(2), v(3), \dots, v(l+1)\} = \{a, a + \cdots, \dots\},$$

where $a > k$ is the smallest element on the right. Since $a > k$, we know that $a' = a$, and hence

$$\{\tilde{v}(1), \tilde{v}(2), \ldots, \tilde{v}(l)\} = \{a', a' + \cdots, \ldots\} = \{a, a + \cdots, \ldots\} \not\leq \{k \leq \cdots\}.$$

This shows that $\tilde{v} \not\leq w_{[k,j]}$.

The remaining case, where neither $k - 1$ nor k is contained in $[i, j]$, can be proved similarly, and it is left as an exercise. This covers all the cases with a Type 3 column with a black lollipop.

Next we consider the case of a Type 3 column with a white lollipop in the kth gap, counting from below in the GP graph, as depicted below:

$$
\begin{array}{ccc}
n-2 & \underline{\hspace{2cm}} & n-1 \\
& \vdots & \\
& \underline{\hspace{2cm}} & \\
k-1 & \circ\!\!-\!\!\!\underline{\hspace{1cm}} & \begin{matrix} k \\ k-1 \end{matrix} \\
& \vdots & \\
1 & \underline{\hspace{2cm}} & 1
\end{array}
$$

By the Type 3 column requirement, we know that

$$\mathscr{R}_0 \overset{s_k s_{k+1} \cdots s_n}{\sim} h(\mathscr{L}_0) \overset{u}{\sim} h(\mathscr{L}^0) \overset{s_n s_{n-1} \cdots s_{n-k}}{\sim} \mathscr{R}^0,$$

where, by assumption, $u \leq w_{[i,j]}$. By Lemma 4.22(2) and a direct computation, we conclude that $\mathscr{R}_0 \overset{v}{\sim} \mathscr{R}^0$ for

$$v \leq \text{Dem}(s_k \ldots s_n \underline{u} s_n \ldots s_{n-k}) \leq \text{Dem}(s_k \ldots s_n \underline{w_{[i,j]}} s_n \ldots s_{n-k}) = \begin{cases} w_{[i,j]} & \text{if } j < k-1, \\ w_{[i,j+1]} & \text{if } i \leq k-1 \leq j, \\ w_{[i+1,j+1]} & \text{if } i \geq k. \end{cases}$$

Lastly, we consider Type 2 columns. By using Lemma 4.22, we compute that, unless we are in one of the two wall shrinking situations

$u \leq w_{[i,j]}$ implies $v \leq w_{[i,j]}$. For the leftmost of the two cases depicted above, we directly have that $u \leq w_{[i,j]}$ implies $v \leq w_{[i,j-1]}$, and, for the rightmost one, $u \leq w_{[i,j]}$ implies $v \leq w_{[i+1,j]}$, as required. $\square$

4.8 Completeness of GP graphs

This brief subsection discusses the concept of *complete* GP graphs, for which the argument we present gives a complete proof of Theorem 1.1. As prefaced in Section 4.2, the factoriality of the coordinate ring $\mathbb{O}(\mathfrak{M}(\Lambda(\mathbb{G}), T))$ is a requirement. Following the results from Section 4.7, we add an additional hypothesis, as follows:

Definition 4.25 A grid plabic graph $\mathbb{G}$ is said to be *complete* if the moduli stack $\mathfrak{M} = \mathfrak{M}(\Lambda(\mathbb{G}), T)$ satisfies the following properties:

- The coordinate ring $\mathbb{O}(\mathfrak{M})$ is a unique factorization domain (UFD).

- For any face $f \subset \mathbb{G}$ with a sugar-free hull $\mathbb{S}_f$, the vanishing locus of the microlocal merodromy A_f is contained in the vanishing loci of A_g for all faces $g \in \mathbb{S}_f$.

These two conditions are technical, and are only trying to capture the most general type of GP graph $\mathbb{G}$ for which the argument works. In practice, if a reasonable example of a $\mathbb{G}$ is given, it is possible to verify the second condition by direct computation (eg using Gröbner bases), whereas the factoriality condition is, to our knowledge, more subtle. That said, as explained in Section 4.2, we have developed combinatorial criteria to ensure the first condition, eg Δ–completeness of $\beta(\mathbb{G})$ or, even more combinatorially, $\mathbb{G}$ being a shuffle graph. In fact, shuffle graphs $\mathbb{G}$ also satisfy the second condition, as can be seen by examining the following combinatorial property:

Definition 4.26 A GP graph $\mathbb{G}$ is said to be $\mathbb{S}$–*complete* if every sugar-free hull of $\mathbb{G}$ can be obtained via some lollipop chain reaction.

By Propositions 4.4 and 4.23, $\mathbb{S}$–complete GP graphs satisfy the second condition in Definition 4.26. By Proposition 2.11, shuffle graphs are $\mathbb{S}$–complete, and therefore complete. The schematic of implications is: shuffle graphs $\mathbb{G} \implies (\beta(\mathbb{G}) \ \Delta$–complete$) + (\mathbb{S}$–complete $\mathbb{G}) \implies$ complete $\mathbb{G}$.

In summary, though Theorem 1.1 is proven for complete grid plabic graphs, there are large classes of $\mathbb{G}$–graphs that can be proven combinatorially to be complete, either because they are shuffle or because $\mathbb{S}$–completeness and Δ–completeness are directly verified. Note that shuffle graphs include all plabic fences, so all open Bott–Samelson varieties at the level of $\mathfrak{M}(\Lambda, T)$, several families of interesting links (such as the twist knots), many braid varieties (eg all 3–stranded ones), and more.

4.9 Proof of the main theorem

In this subsection, we conclude the proof Theorem 1.1. At this stage, we can consider the initial open torus chart in $\mathfrak{M}(\Lambda, T)$ given by $\mathfrak{M}(\mathfrak{w}(\mathbb{G}), T)$, as built in Sections 3 and 4.3, with the candidate cluster $\mathscr{A}$–variables being the microlocal merodromies (constructed in Section 4.6) along a set of initial relative cycles (built in Section 3). Namely, given a GP graph $\mathbb{G}$ and an initial set of relative cycles $\{\eta_1, \ldots, \eta_s\}$ for the pair $(L(\mathbb{G}), T)$ associated to $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$, we have an isomorphism $\mathbb{C}[A_{\eta_1}^{\pm 1}, \ldots, A_{\eta_s}^{\pm 1}] \cong \mathbb{O}(\mathfrak{M}(\mathfrak{w}, T))$, and thus $\{A_{\eta_1}, \ldots, A_{\eta_s}\}$ and their inverses naturally coordinatize the open torus chart $\mathfrak{M}(\mathfrak{w}(\mathbb{G}), T) \subset \mathfrak{M}(\Lambda, T)$. This defines an initial open toric chart U_0, candidate for an initial cluster $\mathscr{A}$–chart, with the quiver $Q(\mathbb{G}, \eta)$ being the intersection quiver associated to the (duals of the) relative cycles $\{\eta_1, \ldots, \eta_s\}$.

We shall now show that the algebra $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ coincides with an upper cluster algebra, along with the remaining items of Theorem 1.1. In geometric terms, the key ingredient for the former claim will be to prove that the initial cluster chart $U_0 \subset \mathfrak{M}(\Lambda, T)$ together with all the once-mutated charts cover the moduli space $\mathfrak{M}(\Lambda, T)$ up to codimension 2, ie

$$U_0 \cup \bigcup_{\eta \text{ mutable}} \mu_\eta(U_0) \overset{\text{up to codim. 2}}{=} \mathfrak{M}(\Lambda, T).$$
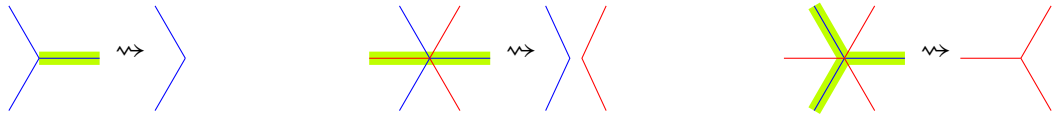
Figure 46: The local models for erasing Y–trees.

By Hartogs's extension theorem, any two normal varieties that differ at most in codimension 2 have the same algebra of regular functions. It thus follows from this codimension 2 isomorphism that $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ is equal to the coordinate ring of the union of the initial chart U_0 and all the once-mutated charts. Then [Berenstein et al. 2005, Corollary 1.9] is used to conclude that the coordinate ring of such a union is an upper cluster algebra.

4.9.1 Erasing Y–trees on weaves and vanishing loci of face merodromies

In order to establish the above covering of $\mathfrak{M}(\Lambda, T)$, up to codimension 2, by U_0 and the once-mutated charts $\mu_\eta(U_0)$, we need to gain understanding of the codimension 1 strata in $\mathfrak{M}(\Lambda, T)$ that appear as vanishing loci of certain microlocal merodromies. These loci can be explicitly described via nonfree weaves that are obtained by erasing Y–cycles in $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$.

First, we begin with the diagrammatic process on the weave that erases Y–trees. Let $\mathfrak{w}$ be a free weave with Y–tree $\gamma \subset \mathfrak{w}$. By definition, the weave $\mathfrak{w}_{\widehat{\gamma}}$ is the weave obtained by erasing $\gamma \subset \mathfrak{w}$ from $\mathfrak{w}$ according to the following local models:

- If γ ends at a trivalent vertex, then we simply erase the weave line contained in γ together with the trivalent vertex, turning the local picture into a single weave line. This is depicted in Figure 46, left.

- If γ goes straight through a hexavalent vertex, then we erase the two weave lines contained in γ and pull the remaining weave lines apart according to their colors, turning the local picture into two weave lines. We draw this in Figure 46, center.

- If γ branches off at a hexavalent vertex, erase the three weave lines contained in γ, turning the local picture into a trivalent weave vertex picture. See Figure 46, left.

Alternatively, it is possible to first shorten the Y–tree to a short I–cycle, and erase the I–cycle, which only requires applying Figure 46, left, twice. The following lemma verifies that this resulting weave is equivalent to the weave obtained by erasing the Y–tree directly:

Lemma 4.27 (Y–tree erasing) *Let $\mathfrak{w}$ be a free weave and $\gamma \subset \mathfrak{w}$ a Y–tree. Let $\mathfrak{w}' \sim \mathfrak{w}$ be a weave obtained from $\mathfrak{w}$ by the double track trick that shortens γ into a short I–cycle (see Proposition 3.5), which we still denote by $\gamma \subset \mathfrak{w}'$. Then there exists a weave equivalence $\mathfrak{w}'_{\widehat{\gamma}} \sim \mathfrak{w}_{\widehat{\gamma}}$. In particular, since $\mathfrak{w}'_{\widehat{\gamma}}$ is not free, neither is $\mathfrak{w}_{\widehat{\gamma}}$.*

Proof Let us start with the short I–cycle γ in $\mathfrak{w}'$. Erasing γ in $\mathfrak{w}'$ leaves two weave lines with a Reeb chord in the middle; see Figure 47, left. Follow the rest of the double tracks and contract these two weave
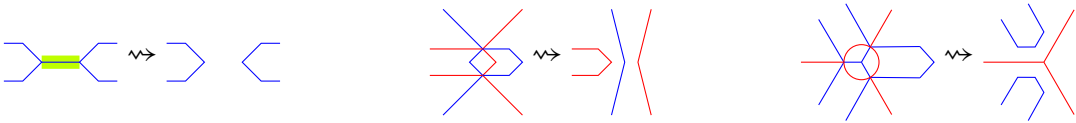
Figure 47: Three types of removals of (pieces of) Y–tree cycles.

lines inductively. At the part where the double track goes straight though, the contracting weave line can be pulled through this part by undoing a candy twist, as depicted in Figure 47, center. At the part where the double tracks branch off, the contracting weave line can be pulled through it by using weave equivalences, thus becoming two contracting weave lines, as illustrated Figure 47, right. In the end, we recover the weave $\mathfrak{w}_{\hat{\gamma}}$, as required. □

Let us now study the vanishing loci of microlocal merodromies A_f associated to faces $f \subset \mathbb{G}$ by using Lemma 4.27. For that, fix a compatible set C of sign curves on the initial weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$ so that the chart $\mathcal{M}_1(\mathfrak{w})$ can be identified with $\mathrm{Loc}_1(L)$. For each relative 1–cycle $\eta \in H_1(L \setminus T, \Lambda \setminus T)$, its microlocal merodromy A_η (Definition 4.10) is well defined and, in particular, we can associate a naive microlocal merodromy A_f with each face f of $\mathbb{G}$. By Corollary 4.12 and Proposition 4.18, these A_f's are regular functions on $\mathfrak{M}(\Lambda, T)$ and they are unique up to multiples of units.

Proposition 4.28 *Let $\mathbb{G}$ be a complete GP graph and $f \subset \mathbb{G}$ a face. Suppose that the lollipop chain reaction initiated from f is complete. Then, for any microlocal merodromy A_f associated with f, the vanishing locus $\{A_f = 0\} \subset \mathfrak{M}(\Lambda(\mathbb{G}), T)$ is nonempty.*

Proof By assumption, f admits a sugar-free hull $\mathbb{S}_f$, which, by Lemma 3.29, gives rise to a Y–tree $\gamma \subset \mathfrak{w}$ in the initial weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$. Apply Proposition 3.5 to γ, making it a short I–cycle, and place this short I–cycle near the end of the original Y–tree, so that it lies inside some Type 1 weave column. If we delete this short I–cycle, we obtain a weave $\mathfrak{w}' = \mathfrak{w}_{\hat{\gamma}}$ whose associated weave surface is immersed, with a single interior Reeb chord at the midpoint of the short I–cycle. We claim that its associated stratum $\mathcal{M}_1(\mathfrak{w}')$ in $\mathfrak{M}(\Lambda(\mathbb{G}), T)$ is nonempty.

To prove this claim, cut the weave $\mathfrak{w}'$ open vertically across the column into two weaves $\mathfrak{w}_1$ and $\mathfrak{w}_2$, so that both $\mathfrak{w}_1$ and $\mathfrak{w}_2$ are free weaves again. By [Jin and Treumann 2017], or [Ekholm et al. 2016; Ng et al. 2020], the two strata $\mathcal{M}_1(\mathfrak{w}_1)$ and $\mathcal{M}_1(\mathfrak{w}_2)$ are nonempty. Now, since this column used to be a Type 1 column, the vertical slice along the cut is a reduced word of $w = s_i w_0$ for some i. Given that any two pairs of flags of relative position w are related to each other by a (nonunique) general linear group element, we use this action to line up the two pairs of flags of relative position w and glue any point in $\mathcal{M}_1(\mathfrak{w}_1)$ with any point in $\mathcal{M}_1(\mathfrak{w}_2)$ and get a point in $\mathcal{M}_1(\mathfrak{w}')$. This shows that $\mathcal{M}_1(\mathfrak{w}')$ is nonempty.

Since $\mathfrak{M}(\mathfrak{w}', T)$ fibers over $\mathcal{M}_1(\mathfrak{w}')$, it follows that $\mathfrak{M}(\mathfrak{w}', T)$ is nonempty as well. Let p be a point in $\mathfrak{M}(\mathfrak{w}', T)$. By Lemma 4.27, the weave $\mathfrak{w}'$ is equivalent to the weave $\mathfrak{w}_{\hat{\gamma}}$. Let η_f be a relative 1–chain associated with f and, without loss of generality, we may assume that η_f is contained in some Type 1

column. By construction, the cross-section of the initial weave $\mathfrak{w}$ at a Type 1 column is always the positive (half-twist) braid Δ, the positive lift of w_0. The erasing of the Y–tree γ turns the cross-sectional positive braid into a positive braid β whose Demazure product satisfies $D(\beta) \leq s_i w_0$ for all i such that the ith gap, counting from below in the GP graph $\mathbb{G}$, is contained in the face f. Moreover, the two flags at the two ends of η_f are of relative position $w \leq D(\beta)$. Thus, we conclude that $w \leq s_i w_0$ and hence $A_f(p) = 0$ by Lemma 4.21. $\qquad\square$

We also establish the converse of Proposition 4.28, ie if the lollipop chain reaction initiated from f is incomplete, then the microlocal merodromy A_f must be a unit in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$.

Proposition 4.29 *Let $\mathbb{G}$ be a GP graph and $f \subset \mathbb{G}$ a face. Then the microlocal merodromy A_f is a unit if and only if the lollipop chain reaction initiated at f is incomplete.*

Proof Indeed, if the lollipop chain reaction initiated from f is complete, then Proposition 4.28 implies that $\{A_f = 0\}$ is nonempty and hence A_f cannot be a unit. For the converse implication, note that the lollipop chain reaction initiated from f being incomplete implies that, during a certain lollipop reaction in the chain, the selection process runs out of faces to select. This is equivalent to saying that the selection process selects an unbounded face g of the GP graph. Now, if there exists a point $p \in \mathfrak{M}(\Lambda, T)$ with $A_f(p) = 0$, then, by Proposition 4.23, $A_g(p) = 0$ as well. But this is impossible because any relative 1–chain η_g associated with g must map to the identity under the projection map $H_1(L \setminus T, \Lambda \setminus T) \to H_1(L, \Lambda)$, and hence A_g is a unit by Corollary 4.12. Therefore, $\{A_f = 0\}$ is empty and A_f is a unit. $\qquad\square$

Finally, let us discuss the ratios of microlocal merodromies that appear when two faces share a sugar-free hull . Recall that the dual basis $\mathfrak{B}^\vee$ of $H_1(L, \Lambda)$ is constructed by starting with the set $\mathfrak{S}(\mathbb{G})$ of initial absolute cycles, which is a linearly independent subset of $H_1(L)$ and is in bijection with the sugar-free hulls of the GP graph $\mathbb{G}$. Each element of $\mathfrak{S}(\mathbb{G})$ is a linear combination of the naive absolute cycles γ_f, which are in bijection with the faces of $\mathbb{G}$. Then we complete $\mathfrak{S}(\mathbb{G})$ to a basis $\mathfrak{B}$ of $H_1(L)$ via a replacement process from bottom to top along the Hasse diagram $\mathscr{H}$ of the sugar-free hulls with respect to inclusion. On the dual side, this replacement process is performed from the top down along the Hasse diagram $\mathscr{H}$, replacing the naive relative cycles η_f one by one and thus obtaining a basis $\mathfrak{B}^\vee$ of $H_1(L, \Lambda)$ dual to $\mathfrak{B}$. By choosing a representative A_f for each naive relative cycle η_f, we construct a microlocal merodromy function A_i for each $i \in \mathfrak{B}^\vee$.

As there can be multiple faces sharing the same sugar-free hull, some faces (naive absolute cycles) are set aside during the replacement process. Suppose $\mathbb{S}_f = \mathbb{S}_g$ for two different faces $f, g \subset \mathbb{G}$ and suppose that we set aside γ_f, while selecting γ_g. Then, on the dual side, the set-aside naive absolute cycle γ_f will correspond to the relative cycle $\eta_f - \eta_g$, which in turn gives rise to the microlocal merodromy function A_f/A_g.

Proposition 4.30 *Let $\mathbb{G}$ be a GP graph and $f, g \subset \mathbb{G}$ two faces with equal sugar-free hulls. Then the microlocal merodromy A_f/A_g corresponding to a set-aside naive absolute cycle γ_f is a unit.*

Proof Since the faces f and g have the same sugar-free hull, the lollipop chain reaction initiated from one of them must contain the other. Therefore, by Proposition 4.23, $A_f(p) = 0$ if and only if $A_g(p) = 0$ for any $p \in \mathfrak{M}(\Lambda, T)$. Since $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ is a UFD, this implies that A_f and A_g are associates of each other, and thus A_f/A_g is a unit. □

4.9.2 Rank of exchange matrix and mutation formulae for Lagrangian surgeries Recall the notation $U_0 = \mathfrak{M}(\mathfrak{w}, T) \subset \mathfrak{M}(\Lambda, T)$ for the open toric chart associated to the (Lagrangian filling for the) weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$. Let us denote the naive microlocal monodromies by $\{A_f\}$, where f runs through the faces of $\mathbb{G}$, and the microlocal monodromies associated with marked points by $\{A_t\}_{t \in T}$. The microlocal merodromies associated with the dual basis $\mathfrak{B}^\vee$ will be denoted by $\{A_i\}$, where $i \in [1, b_1(L)]$. By construction,

$$\mathbb{O}(U_0) = \mathbb{C}[A_{f_1}^{\pm 1}, \ldots, A_{f_{b_1(L)}}^{\pm 1}, A_t^{\pm 1}] = \mathbb{C}[A_1^{\pm 1}, \ldots, A_{b_1(L)}^{\pm 1}, A_t^{\pm 1}],$$

where the variable t runs through the set of marked points T. Let us also fix $\widetilde{\mathfrak{B}}$ to be a completion of the basis $\mathfrak{B} \subset H_1(L)$ to a basis of $H_1(L, T)$. Then, by possibly adding relative 1–chains $\{\xi_t\}_{t \in T}$, we can modify elements of $\mathfrak{B}^\vee$ so that $\mathfrak{B}^\vee \sqcup \{\xi_t\}_{t \in T}$ becomes a dual basis of $\widetilde{\mathfrak{B}}$. In this modification, each microlocal merodromy A_i is multiplied by some Laurent monomial in the A_t's. To ease notation, we rename the microlocal merodromies A_i to include these Laurent monomials as well. Since $\widetilde{\mathfrak{B}}$ and $\widetilde{\mathfrak{B}}^\vee$ are dual bases, there is a natural bijection between them, and we will use them interchangeably as index sets for microlocal monodromies and microlocal merodromies.

Now, the intersection form $\{\cdot, \cdot\}$ on $H_1(L)$ can be extended to a skew-symmetric form on $H_1(L, T)$ by imposing a half-integer value for intersections at T:

$$\diagup\!\!\!\!\!\nwarrow\!\!\!* = \frac{1}{2} \diagup\!\!\!\!\!* \diagdown$$

With respect to the basis $\widetilde{\mathfrak{B}}$, the intersection form on $H_1(L, T)$ is then encoded by a $\widetilde{\mathfrak{B}} \times \widetilde{\mathfrak{B}}$ skew-symmetric matrix ϵ, where

$$\epsilon_{ij} = \{\gamma_i, \gamma_j\}$$

for any pair $\gamma_i, \gamma_j \in \widetilde{\mathfrak{B}}$. For any absolute 1–cycle γ in $H_1(L)$, Section 4.4 constructs the microlocal monodromy function ψ_γ on $\mathcal{M}_1(\mathfrak{w})$. After the correction by sign curves, we have $X_i := \psi_{\gamma_i}$, and the collection $\{X_i\}_{i \in \mathfrak{B}}$ are our candidates for the initial cluster $\mathcal{X}$–variables.

Let $p: \mathfrak{M}(\Lambda, T) \to \mathcal{M}_1(\Lambda)$ be the forgetful map. Then, by restricting to the respective tori supported on the initial weave $\mathfrak{w}$, we also obtain $p: \mathfrak{M}(\mathfrak{w}, T) \to \mathcal{M}_1(\mathfrak{w})$.

Proposition 4.31 *Consider the forgetful map $p: \mathfrak{M}(\mathfrak{w}, T) \to \mathcal{M}_1(\mathfrak{w})$. For an initial absolute cycle $\gamma_i \in \mathfrak{S}(\mathbb{G})$,*

$$p^*(X_i) = \prod_{j \in \widetilde{\mathfrak{B}}} A_j^{\epsilon_{ij}}.$$
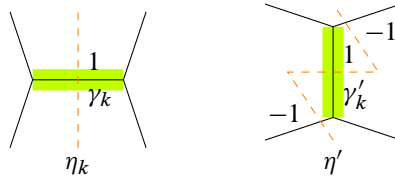
Figure 48: Mutation of initial relative cycles.

Proof Let us denote the relative 1–chain dual to $\gamma_i \in \widetilde{\mathfrak{B}}$ by η_i. It suffices to prove that

$$\gamma_i = \sum_j \epsilon_{ij} \eta_j$$

under the inclusion map $H_1(L) \cong H_1(L \setminus T) \hookrightarrow H_1(L \setminus T, \Lambda \setminus T)$. Note that, since we have at least one marked point per link component in Λ, we can lift elements from $H_1(L, \Lambda)$ to $H_1(L, T)$. Now, for any element $\theta = \sum_{k \in \mathfrak{F}} c_k \gamma_k \in H_1(L, T)$ that is a lift of a relative 1–cycle in $H_1(L, \Lambda)$,

$$\left\langle \sum_j \epsilon_{ij} \eta_j, \theta \right\rangle = \left\langle \sum_j \epsilon_{ij} \eta_j, \sum_k c_k \gamma_k \right\rangle = \sum_j \epsilon_{ik} c_k = \langle \gamma, \theta \rangle.$$

Since the intersection form is nondegenerate on the tensor product $H_1(L) \otimes H_1(L, \Lambda)$, it indeed follows that $\gamma = \sum_j \epsilon_{ij} \eta_j$. □

Corollary 4.32 *The rectangular exchange matrix $\epsilon|_{\mathfrak{S}(\mathbb{G}) \times \widetilde{\mathfrak{B}}}$ is full-ranked.*

Proof Since $\epsilon|_{\mathfrak{S}(\mathbb{G}) \times \widetilde{\mathfrak{B}}}$ is a submatrix of $\epsilon|_{\mathfrak{B} \times \widetilde{\mathfrak{B}}}$, it suffices to prove that $\epsilon|_{\mathfrak{B} \times \widetilde{\mathfrak{B}}}$ is full-ranked. The latter follows from the surjectivity of the map $p: \mathfrak{M}(\mathfrak{w}, T) \to \mathcal{M}_1(\mathfrak{w})$. □

Let us now focus on the effect that a Lagrangian surgery, in the form of weave mutation, has on microlocal merodromies. For that we need to understand how relative cycles change under such an operation, as follows. Let γ_k be an initial absolute cycle which, by Lemma 3.29, we represent as a Y–tree on the initial weave $\mathfrak{w} = \mathfrak{w}(\mathbb{G})$. By Proposition 3.5, there exists a weave equivalence that isotopes γ_k to a short I–cycle. In this weave equivalence, the dual basis element η_k of γ_k must also be isotoped to a curve that cuts through this short I–cycle in the middle. Thus, near the short I–cycle γ_k, the local model is the one depicted in Figure 48, left. Note that each of the four weave lines extending out of this local picture may be part of multiple initial absolute cycles. However, we may assume without loss of generality that all other initial relative cycles are outside this local picture.

By performing a weave mutation at γ_k, we obtain a new weave $\mathfrak{w}_k$, which is mostly identical to $\mathfrak{w}$ except the local picture is replaced by Figure 48. Note that the initial absolute cycle γ_k in $\mathfrak{w}$ is replaced by the new absolute cycle γ'_k in $\mathfrak{w}_k$.

In the local model in Figure 48, we have also drawn a relative 1–chain η', which connects to the rest of $-\eta_k$ outside of the local picture. But this relative 1–chain η' is not the correct replacement for η_k after

the surgery, because, in addition to having intersection 1 with γ_k', the new relative cycle η_k' also needs to have trivial intersection with all other absolute cycles in $\mathfrak{B}$. At this stage, η' could possibly have nontrivial intersections with absolute cycles that come into the local picture from the northeast and the southeast. Thus, the correct replacement for the relative cycle η_k after the weave mutation is the linear combination

$$(4\text{-}11) \qquad \eta_k' = \eta' + \sum_{j \in \widetilde{\mathfrak{B}}} [-\epsilon_{kj}]_+ \eta_j.$$

This explains how to keep track of relative cycles after a weave mutation, and thus a Lagrangian surgery in our context. Note that the moduli space $\mathfrak{M}(\mathfrak{w}_k, T)$ also defines an open toric chart $\mathfrak{M}(\mathfrak{w}_k, T) \subset \mathfrak{M}(\Lambda, T)$, as $\mathfrak{w}_k$ defines an embedded exact Lagrangian filling as well. Let us denote this chart, where we have performed a Lagrangian surgery at the kth disk of the $\mathbb{L}$–compressing system, by $U_k \subset \mathfrak{M}(\Lambda, T)$, and denote the microlocal merodromy associated with the relative cycle η_k' by A_k'. In order to understand the change of the microlocal merodromies under surgery, we have the following result:

Proposition 4.33 *At any point $u \in U_0 \cap U_k$,*

$$A_k' = A_k^{-1}(1 + p^*(X_k)) \prod_{j \in \widetilde{\mathfrak{B}}} A_j^{[-\epsilon_{kj}]_+},$$

where $p \colon U_0 \to \mathcal{M}_1(\mathfrak{w})$ is the forgetful map restricted to $U_0 \cap U_k$.

Proof It suffices to prove that $A_{\eta'} = A_k^{-1}(1 + p^*(X_k))$. Since γ_k is a short I–cycle, we can assume that the four neighboring flags are four lines l_e, l_s, l_w and l_n in $\mathbb{C}^2$. Let v_i be a nonzero vector in each line l_i and let det denote the dual of the nonzero 2–form associated with $\mathbb{C}^2$. Following the crossing-value formula, we obtain

$$A_k = \det(v_s \wedge v_n), \quad A_{\eta'} = \frac{\det(v_e \wedge v_w)}{\det(v_n \wedge v_e)\det(v_s \wedge v_w)}.$$

Therefore, the product can be computed as

$$A_k A_{\eta'} = \frac{\det(v_s \wedge v_n)\det(v_e \wedge v_w)}{\det(v_n \wedge v_e)\det(v_s \wedge v_w)} = \frac{\det(v_n \wedge v_e)\det(v_s \wedge v_w) + \det(v_n \wedge v_w)\det(v_e \wedge v_s)}{\det(v_n \wedge v_e)\det(v_s \wedge v_w)}$$
$$= 1 + p^*(X_k). \qquad \square$$

Remark 4.34 Instead of the relative 1–chain η' depicted in Figure 48, right, we could have chosen η' to have support the other zigzag $\diagdown$ with appropriate intersection numbers. In this other choice, equation (4-11) would need to be modified to $\eta_k' = \eta' + \sum_{j \in \widetilde{\mathfrak{B}}} [\epsilon_{kj}]_+ \eta_j$ and the equation in Proposition 4.33 would be modified to $A_k' = A_k^{-1}(1 + p^*(X_k^{-1})) \prod_{j \in \widetilde{\mathfrak{B}}} A_j^{[\epsilon_{kj}]_+}$. These compatible changes of signs $\epsilon_{kj} \to -\epsilon_{kj}$ define the chiral dual cluster structure on the same variety, which is the cluster structure defined by the opposite quiver. This chiral dual is discussed in [Fock and Goncharov 2009, Section 1.2]. Since the quiver is given by the intersection pairing between absolute cycles, the choice in Figure 48 is, in a sense, naturally dictated by the chosen orientation on the filling $L(\mathfrak{w})$.

Propositions 4.33 and 4.31 yield the desired cluster $\mathcal{A}$–mutation formula for the microlocal merodromies under Lagrangian surgeries on the set of initial $\mathbb{L}$–compressing disks:

Corollary 4.35 *Let $\mathbb{G}$ be a GP graph, $\{\eta_1, \ldots, \eta_s\}$ the set of naive relative cycles and $\{A_i\}$ the associated set of naive microlocal merodromies. Consider the microlocal merodromy A'_k along the relative 1–chain η'_k obtained from η_k by weave mutation at the dual 1–cycle γ_k. Then*

$$A'_k = \frac{\prod_{j \in \mathfrak{B}} A_j^{[\epsilon_{kj}]+} + \prod_{j \in \mathfrak{B}} A_j^{[-\epsilon_{kj}]+}}{A_k}.$$

4.9.3 Regularity of initial microlocal merodromies By Proposition 4.18, the naive microlocal merodromies A_f are regular functions on the moduli space $\mathfrak{M}(\Lambda, T)$. However, since the adjusted microlocal merodromies $\{A_i\}_{i \in \widetilde{\mathfrak{B}}}$ corresponding to the initial basis are ratios of the naive microlocal merodromies, the initial merodromies $\{A_i\}$ are only rational functions a priori. Our next goal is to prove that, for all $i \in \mathfrak{B}$, the A_i are actually global regular functions, and that they are either irreducible if $i \in \mathfrak{S}(\mathbb{G})$, or units otherwise. Let us start with the following lemma:

Lemma 4.36 *Let $U_0 \subset \mathfrak{M}(\Lambda, T)$ be the initial open toric chart and f a unit in $\mathbb{O}(U_0)$ (resp. $\mathbb{O}(U_k)$). Suppose that $f = gh$ in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ for some $g, h \in \mathbb{O}(\mathfrak{M}(\Lambda, T))$. Then g and h are also units in $\mathbb{O}(U_0)$ (resp. $\mathbb{O}(U_k)$).*

Proof Indeed, if $f = gh$ in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$, then $f = gh$ in $\mathbb{O}(U_0)$ as well, and, if f is a Laurent monomial in $\mathbb{O}(U_0)$, then each of g and h must be a Laurent monomial, too. The proof for the case where f is a unit in $\mathbb{O}(U_k)$ is analogous.   □

We first show that the initial merodromies are irreducible assuming they are regular functions:

Lemma 4.37 *Let $\mathbb{G}$ be a GP graph, $\gamma_k \in \mathfrak{S}(\mathbb{G})$ an initial absolute cycle and consider $A_k : \mathfrak{M}(\Lambda, T) \dashrightarrow \mathbb{C}$ an associated microlocal merodromy. Suppose that A_k is a regular function, ie an element of $\mathbb{O}(\mathfrak{M}(\Lambda, T))$. Then A_k is irreducible.*

Proof Suppose $A_k = gh$ in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ with neither g nor h being a unit. Then Lemma 4.36 implies that g and h must be Laurent monomials , and hence can be expressed as $\prod_{i \in \widetilde{\mathfrak{B}}} A_i^{m_i}$ and $\prod_{i \in \widetilde{\mathfrak{B}}} A_i^{n_i}$, respectively, up to a multiple of units in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$. Since $A_k = \prod_j A_i^{m_i + n_i}$, then at least one of m_k and n_k must be positive. Without loss of generality, let us assume that $m_k > 0$. Then, since h is not a unit, there must be some $j \neq k$ such that A_j is not a unit and $n_j > 0$. If A_j is not a unit, then, by Proposition 4.29, j must correspond to an initial absolute cycle γ_j. By mutating along the initial absolute cycle γ_j, we obtain a new weave $\mathfrak{w}_j$. By Lemma 4.36, h is also a Laurent monomial in the new chart $\mathbb{O}(U_j)$ associated to $\mathfrak{w}_j$ and hence we can write g and h as

$$g = A_j'^{p_j} \prod_{i \neq j} A_i^{p_i} \quad \text{and} \quad h = A_j'^{q_j} \prod_{i \neq j} A_i^{q_i}.$$

Note that, since $A_k = gh$, we must have $p_j + q_j = 0$. If $p_j = q_j = 0$, then we have a contradiction because $\prod_{i \neq j} A_i^{q_i} = h = \prod_i A_i^{n_i}$ with $n_j > 0$. That said, if p_j and q_j are nonzero, then one of them must be positive; suppose $p_j > 0$. By Corollary 4.35, $A_j' = M_1 + M_2$, where M_1 and M_2 are two algebraically independent Laurent monomials in $\{A_i\}_{i \in \widetilde{\mathfrak{B}}}$, up to units. It then follows that

$$g = (M_1 + M_2)^{p_j} \prod_{i \neq j} A_i^{p_i},$$

which shows that g is not a Laurent monomial in $\{A_i\}_{i \in \widetilde{\mathfrak{B}}}$. This is again a contradiction, and therefore A_k must be an irreducible element in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$. $\square$

We are ready to conclude regularity, and thus irreducibility, of initial merodromies:

Proposition 4.38 *Let $\mathbb{G}$ be a GP graph, $\gamma_k \in \mathfrak{S}(\mathbb{G})$ an initial absolute cycle and A_k an associated microlocal merodromy. Then A_k is a regular function and an irreducible element in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$.*

Proof By Lemma 4.37, it suffices to prove that A_k is a regular function. We proceed by induction from top down along the Hasse diagram $\mathcal{H}$; recall that vertices of $\mathcal{H}$ are sugar-free hulls and hence they are naturally indexed by the set of initial absolute cycles $\mathfrak{S}(\mathbb{G})$. For the base case, suppose k is a maximal vertex in the Hasse diagram. Then $A_k = A_f$ for some naive relative cycle η_f. Then Proposition 4.18 implies that $A_k = A_f$ is a regular function, as required. Inductively, suppose, for all $i > k$ in the Hasse diagram $\mathcal{H}$, A_i is a regular function on $\mathfrak{M}(\Lambda, T)$. By Lemma 4.37, A_i are irreducible elements in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ as well. Let f_i be the face selected for each vertex i of $\mathcal{H}$. Then, for each vertex i of $\mathcal{H}$,

$$A_{f_i} = \prod_{j \geq i} A_j.$$

In particular, if $i > k$, then the above is the unique factorization of the naive microlocal merodromies A_{f_i} in $\mathbb{O}(\mathfrak{M}(\Lambda, T))$, up to multiple of units. This also implies that the irreducible elements $\{A_i\}_{i > k}$ are not associates of each other because A_{f_i} are not units by Corollary 4.12.

In addition to the above, if $i > k$, then f_i is contained in the sugar-free hull $\mathbb{S}_{f_i} = \mathbb{S}_i$. Proposition 4.23 implies $\{A_{f_i} = 0\} \subset \{A_{f_k} = 0\}$, but we obtain the inclusion $\{A_i = 0\} \subset \{A_{f_i} = 0\}$ as well because A_i is an irreducible factor of A_{f_i}. Therefore, $\{A_i = 0\} \subset \{A_{f_k} = 0\}$, which is equivalent to A_{f_k} being divisible by A_i for all $i > k$. Since A_i are distinct irreducible elements of $\mathbb{O}(\mathfrak{M}(\Lambda, T))$, it follows that the quotient

$$A_k = A_{f_k} \prod_{i > k} A_i^{-1}$$

is also a regular function on $\mathfrak{M}(\Lambda, T)$ as well. The induction is now complete. $\square$

4.9.4 Conclusion of the argument We finalize the proof of the covering of $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ by the initial and adjacent charts, up to codimension 2. For each initial absolute cycle $\gamma_k \in \mathfrak{S}(\mathbb{G})$, we denote the vanishing locus of the associated microlocal merodromy by $D_k := \{A_k = 0\} \subset \mathfrak{M}(\Lambda, T)$. Since A_k is an irreducible element of $\mathbb{O}(\mathfrak{M}(\Lambda, T))$, D_k is irreducible as a codimension 1 subvariety in $\mathfrak{M}(\Lambda, T)$.

Proposition 4.39 *Let $\mathbb{G}$ be a GP graph, $\gamma_k \in \mathfrak{S}(\mathbb{G})$ the initial absolute cycles, $U_k \subset \mathfrak{M}(\Lambda, T)$ the open torus chart associated to the Lagrangian surgery of $L(\mathfrak{w}(\mathbb{G}))$ at γ_k, and $D_k \subset \mathfrak{M}(\Lambda, T)$ the vanishing locus of its associated microlocal merodromy. Then the intersection $U_k \cap D_k \subset \mathfrak{M}(\Lambda, T)$ is a nonempty open subset of the vanishing locus D_k.*

Proof It suffices to prove that $U_k \cap D_k$ is nonempty. Similar to the proof of Proposition 4.28, we apply Proposition 3.5 to move γ_k to a short I–cycle near the end of the original Y–tree, so that it lies inside some Type 1 weave column. By deleting this short I–cycle, we obtain a weave $\mathfrak{w}'$ whose moduli space $\mathfrak{M}(\mathfrak{w}', T)$ is a subset of D_k. It suffices to show that $\mathfrak{M}(\mathfrak{w}', T) \cap U_k \neq \varnothing$, but this clear: For instance, in the case where the weave looks like the one on the left in
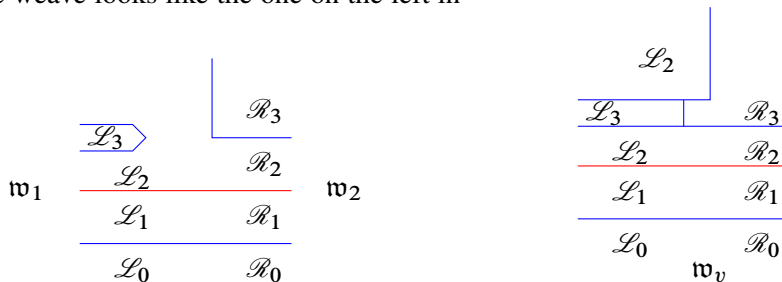
we first fix a point in $\mathcal{M}_1(\mathfrak{w}_1)$ and then, based on the flags $\mathscr{L}_0$, $\mathscr{L}_1$, $\mathscr{L}_2$ and $\mathscr{L}_3$, we choose a point in $\mathcal{M}_1(\mathfrak{w}_2)$ with flags $\mathscr{R}_0 = \mathscr{L}_0$, $\mathscr{R}_1 = \mathscr{L}_1$ and $\mathscr{R}_2 = \mathscr{L}_2$, but $\mathscr{R}_3 \neq \mathscr{L}_3$, and then glue them together. This gives a point in $\mathcal{M}_1(\mathfrak{w}')$ which is also in $\mathcal{M}_1(\mathfrak{w}_k)$, where $\mathfrak{w}_k$ is the mutated weave, which is also shown on the right above. $\square$

At this stage, the covering property, up to codimension 2, readily follows:

Proposition 4.40 *Let $\mathbb{G}$ be a GP graph, $\gamma_k \in \mathfrak{S}(\mathbb{G})$ the initial absolute cycles, and $U_k \subset \mathfrak{M}(\Lambda, T)$ the open torus charts associated to the Lagrangian surgery of $L(\mathfrak{w}(\mathbb{G}))$ at each γ_k, where U_0 is the initial chart associated to $L(\mathfrak{w}(\mathbb{G}))$. Then*

$$\operatorname{codim}\left(U_0 \cup \bigcup_{k \in \mathfrak{S}(\mathbb{G})} U_k \right) \geq 2,$$

ie the union of U_0 and all the adjacent charts U_k covers $\mathfrak{M}(\Lambda, T)$ up to codimension 2.

Proof By Proposition 4.39, the intersection $U_j \cap D_j$ is open in D_j for all j, and thus $\operatorname{codim}(D_j \cap U_j^c) \geq 2$ for each j. Thus, $\operatorname{codim}\left(U_0 \cup \bigcup_k U_k \right) \geq 2$ by the inclusions

$$\left(U_0 \cup \bigcup_k U_k \right)^c = U_0^c \cap \bigcap_k U_k^c = \left(\bigcup_j D_j \right) \cap \bigcap_k U_k^c = \bigcup_j \left(D_j \cap \bigcap_k U_k^c \right) \subset \bigcup_j (D_j \cap U_j^c). \quad \square$$

Theorem 1.1 and Corollary 1.2 are now concluded as follows:

Theorem 4.41 *Let $\mathbb{G}$ be a complete GP graph. The coordinate ring of regular functions $\mathbb{O}\big(\mathfrak{M}(\Lambda(\mathbb{G}), T)\big)$ has the structure of an upper cluster algebra.*

Proof Consider the open subset $U_0 \cup \bigcup_{k \in \mathfrak{S}(\mathbb{G})} U_k \subset \mathfrak{M}(\Lambda, T)$. Proposition 4.40 shows the equality of coordinate rings $\mathbb{O}(\mathfrak{M}(\Lambda, T)) = \mathbb{O}(U_0 \cup \bigcup_{k \in \mathfrak{S}(\mathbb{G})} U_k)$. Corollary 4.35 implies that $\mathbb{O}(U_0 \cup \bigcup_{k \in \mathfrak{S}(\mathbb{G})} U_k)$ is an upper bound of a cluster algebra. In addition, since T has at least one marked point per link component, Corollary 4.32 shows that the rectangular exchange matrix $\epsilon|_{\mathfrak{S}(\mathbb{G}) \times \widetilde{\mathfrak{B}}}$ is full-ranked. Then [Berenstein et al. 2005, Corollary 1.9] implies that this upper bound coincides with its upper cluster algebra and therefore we conclude that $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ is an upper cluster algebra. □

Corollary 4.42 *Let $\mathbb{G}$ be a complete GP graph. Then $\mathbb{O}(\mathcal{M}_1(\Lambda(\mathbb{G})))$ has the structure of a cluster Poisson algebra.*

Proof Let us temporarily denote the cluster $\mathscr{A}$–variety defined by the $\widetilde{\mathfrak{B}} \times \widetilde{\mathfrak{B}}$ exchange matrix ϵ by $\mathscr{A}$ and denote the cluster $\mathscr{X}$–variety associated with the submatrix $\epsilon|_{B \times B}$ by $\mathscr{X}$. Since $\epsilon|_{B \times \widetilde{B}}$ is full-ranked, which follows from the surjectivity of $p \colon \mathfrak{M}(\Lambda, T) \to \mathcal{M}_1(\Lambda)$, the cluster-theoretical map $p \colon \mathscr{A} \to \mathscr{X}$ is also surjective. Both $\mathbb{O}(\mathscr{A})$ and $\mathbb{O}(\mathscr{X})$ are intersections of Laurent polynomial rings, and thus a rational function f on $\mathscr{X}$ is regular if and only if $p^*(f)$ is regular on $\mathscr{A}$; see [Shen and Weng 2020, Lemma A.1]. That said, given that $p \colon \mathfrak{M}(\Lambda, T) \to \mathcal{M}_1(\Lambda)$ is surjective, a rational function on $\mathcal{M}_1(\Lambda)$ is regular if and only if $p^*(g)$ is regular on $\mathfrak{M}(\Lambda, T)$. Now consider the commutative diagram

$$
\begin{array}{ccc}
\mathfrak{M}(\Lambda, T) & \xrightarrow[\cong]{\alpha} & \mathscr{A} \\
\downarrow{\scriptstyle p} & & \downarrow{\scriptstyle p} \\
\mathcal{M}_1(\Lambda) & \xrightarrow[\chi]{} & \mathscr{X}
\end{array}
$$

Both horizontal maps are birational because $\mathfrak{M}(\Lambda, T)$ (resp. $\mathcal{M}_1(\Lambda)$) and $\mathscr{A}$ (resp. $\mathscr{X}$) share an open torus chart $\mathfrak{M}(\mathfrak{w}, T)$ (resp. $\mathcal{M}_1(\mathfrak{w})$). In addition, Theorem 4.41 implies that the top map induces an isomorphism between $\mathbb{O}(\mathfrak{M}(\Lambda, T))$ and $\mathbb{O}(\mathscr{A})$. Now, given a regular function $f \in \mathbb{O}(\mathscr{X})$, the pullback $\chi^*(f)$ is a rational function on $\mathcal{M}_1(\Lambda)$ by birationality; but, since $p^* \circ \chi^*(f) = \alpha^* \circ p^*(f)$ is regular on $\mathfrak{M}(\Lambda, T)$, it follows that $\chi^*(f)$ is regular on $\mathcal{M}_1(\Lambda)$. Conversely, if we are given a regular function $f \in \mathbb{O}(\mathcal{M}_1(\Lambda))$, we know that $(\chi^{-1})^*(f)$ is a rational function on $\mathscr{X}$ by birationality; but, since $p^* \circ (\chi^{-1})^*(f) = (\alpha^{-1})^* \circ p^*(f)$ is regular on $\mathscr{A}$, it follows that $(\chi^{-1})^*(f)$ is regular on $\mathscr{X}$ as well. Therefore, we conclude that χ induces an algebra isomorphism between $\chi^* \colon \mathbb{O}(\mathscr{X}) \to \mathbb{O}(\mathcal{M}_1(\Lambda))$, and hence $\mathbb{O}(\mathcal{M}_1(\Lambda))$ is a cluster Poisson algebra. □

5 Cluster DT transformations for shuffle graphs

The cluster Donaldson–Thomas (DT) transformation is a cluster variety automorphism that manifests the Donaldson–Thomas invariants of a 3D Calabi–Yau category associated with the cluster ensemble [Kontsevich and Soibelman 2010; Keller 2017; Goncharov and Shen 2018]. In this section we prove Corollary 1.3,

ie we focus on the cluster varieties $\mathcal{M}_1(\Lambda)$ associated with shuffle graphs and in particular show that their cluster DT transformation is the composition of a Legendrian isotopy and a contactomorphism of $(\mathbb{R}^3, \xi_{st})$.

5.1 Initial quivers of shuffle graphs

Let us first prove features of the initial quivers associated with shuffle graphs. From now onward, we assume without loss of generality that shuffle graphs have all vertical edges with a black vertex on top.
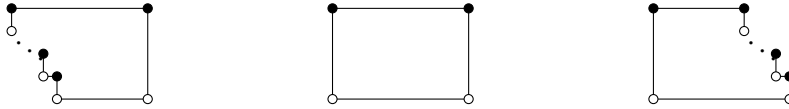
Proposition 5.1 *Let $\mathbb{G}$ be a shuffle graph and $f \subset \mathbb{G}$ a face. Then the sugar-free hull $\mathbb{S}_f$ can have a staircase pattern on at most one of its sides.*

Proof If a sugar-free hull has staircase patterns on more than one of its sides, then somewhere in this sugar-free hull we must have two opposing staircases that look like
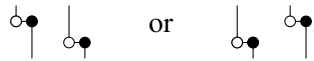
In either case, the horizontal lines containing the horizontal edges above violate Definition 2.8. □

Corollary 5.2 *Let $\mathbb{G}$ be a shuffle graph. Then all its sugar-free hulls must be one of the three shapes*

Proposition 5.3 *Let $\mathbb{G}$ be a shuffle graph. If $\mathbb{S}_f$ and $\mathbb{S}_g$ are two sugar-free hulls and $\mathbb{S}_f \subset \mathbb{S}_g$, then there is no arrow between their corresponding quiver vertices $Q(\mathbb{G})$, ie $\langle \partial \mathbb{S}_f, \partial \mathbb{S}_g \rangle = 0$.*

Proof If $\mathbb{S}_f$ and $\mathbb{S}_g$ do not share boundaries, then $\langle \partial \mathbb{S}_f, \partial \mathbb{S}_g \rangle = 0$. If $(\partial \mathbb{S}_f) \cap (\partial \mathbb{S}_g) \neq \varnothing$, then, based on their possible shapes listed in Corollary 5.2, we see that $(\partial \mathbb{S}_f) \cap (\partial \mathbb{S}_g)$ must be the union of a consecutive sequence of edges. By going over all possibilities of having opposite colors at the two endpoints, we deduce that each possibility will always cut $\mathbb{S}_g$ into smaller sugar-free regions, making $\mathbb{S}_g$ no longer a sugar-free hull. Thus, the two endpoints of this union must be of the same color. Note that the pairing $\langle \partial \mathbb{S}_f, \partial \mathbb{S}_g \rangle$ can be computed by summing over contributions from the bipartite edges in $(\partial \mathbb{S}_f) \setminus (\partial \mathbb{S}_g)$: since the two endpoints of $(\partial \mathbb{S}_f) \setminus (\partial \mathbb{S}_g)$ are the same as the two endpoints of $(\partial \mathbb{S}_f) \cap (\partial \mathbb{S}_g)$, we can conclude that the contributions from the bipartite edges must cancel each other out, leaving $\langle \partial \mathbb{S}_f, \partial \mathbb{S}_g \rangle = 0$ as a result. □

By Definition 2.8, a shuffle graph $\mathbb{G}$ with n horizontal lines is equipped with a permutation $\sigma \in S_n$. Based on the permutation σ, we decompose the vertex set of the initial quiver $Q(\mathbb{G})$ as follows:
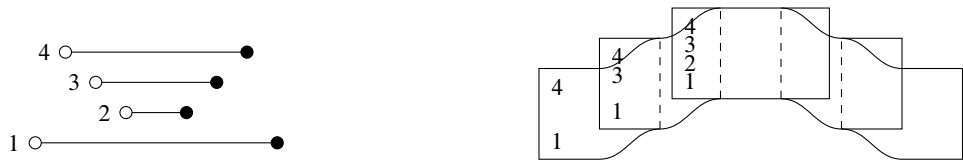
Figure 49: Left: the relative lengths of horizontal lines in a shuffle graph $\mathbb{G}$ associated with σ. Right: the branching of the quiver $Q(\mathbb{G})$; the collection of numbers on each branching records the horizontal lines that define the levels for quiver vertices on that branch.

Definition 5.4 For each integer m with $1 \leq m < n$, we define $\sigma^{-1}[m, n]$ to be the preimage of the $(n-m+1)$–element set $[m, n]$. We order elements in $\sigma^{-1}[m, n]$ according to the ordinary linear order on natural numbers. We say that (i, j) form a *level* if $i < j$ in $\sigma^{-1}[m, n]$ for some m and there is no $k \in \sigma^{-1}(m, n)$ such that $i < k < j$. We say a quiver vertex is on level (i, j) if its corresponding sugar-free hull is sandwiched between the ith and the jth horizontal lines.

If $\sigma^{-1}(m) = k$ and there exists $i < j$ in $\sigma^{-1}[m + 1, n]$ such that $i < k < j$, then there can be sugar-free hulls on level (i, j) containing sugar-free hulls on levels (i, k) and (k, j). It is possible to visualize this phenomenon as a branching on the quiver $Q(\mathbb{G})$: the *main branch* contains sugar-free hulls on levels (i, k) and (k, j) and the *side branch* contains sugar-free hulls on level (i, j). See Figures 49, right, and 50. Note that such a branching may happen multiple times, with the side branch of the former branching becoming the main branch of the next. Figure 50 illustrates this for a shuffle graph associated with the permutation $\sigma = [4\,1\,2\,3]$, with two branchings on each side.

5.2 Reflection moves

In order to geometrically construct the DT transformations for general shuffle graphs, we now generalize the left and right reflection moves introduced in [Shen and Weng 2021].

Consider a Type 2 weave column with an outgoing weave line s_i on one side (top or bottom). By using weave equivalences, we can extend this outgoing weave line inward, penetrating through the weave column and forming a trivalent weave vertex on the other side with color s_{n-i}. If, in addition, either of
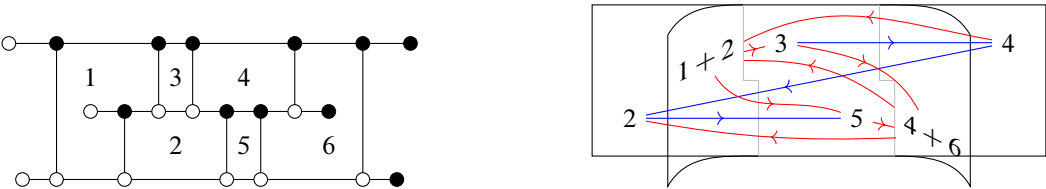
Figure 50: Example of the branching phenomenon of the quiver of a shuffle graph. The blue arrows lie on the main branch (the plabic fence part). The red arrows go between the side branches and the main branch.
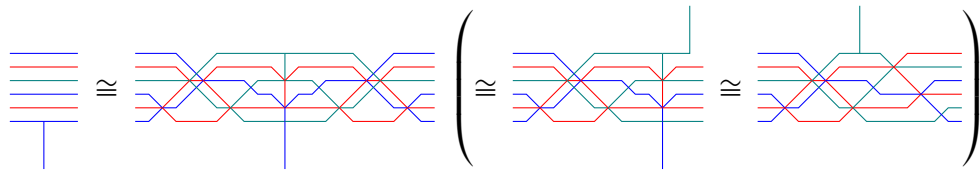
Figure 51: The first two pictures are an example of the local penetration move. If there is an additional weave equivalence that turns the 2nd picture into the 3rd one with an outgoing weave line on the top, then we can use it to turn the 3rd one back to a Type 2 weave column again.

the two horizontal weave lines incident to the new trivalent weave vertex happens to be outgoing as well, then we can homotope the weave locally so that the local picture becomes a Type 2 weave column with an outgoing weave line s_{n-i} on the other side.

These weave equivalences are more general than reflection moves for rainbow closures in [loc. cit.]. In our upcoming construction of cluster DT transformations, we will apply this weave equivalence to lollipops. For example, suppose b is a black lollipop on the ith horizontal line in a grid plabic graph $\mathbb{G}$. Take the first vertical edge e' with a black vertex on the ith horizontal line as we search from right to left starting from b. Suppose the other vertex of e' lies on the jth horizontal line whose right endpoint lies to the right of b, and suppose there are no more vertical edges (of either pattern) between the ith and j^{th} horizontal lines to the right of e'. Then the reflection move can be used to turn e' into its opposite pattern; a side-effect is that this move would also switch the portions of the ith and j^{th} horizontal lines on the right side of e', resulting in a possibly nonplanar bicolor graph. Figure 52 gives an example of such a move done on a plabic graph with three horizontal lines. A similar move can be applied to a white lollipop w, and the vertical edge is found by scanning rightward from w.

Since such a move can potentially destroy planarity to the right of edge e', sugar-free hulls no longer make sense there. However, if the part of the quiver corresponding to the region on the right of edge e' does not get involved in the current iterative step, then this does not affect the construction of the cluster DT transformations. Moreover, the reflection move can enable us to pass a vertical edge through an
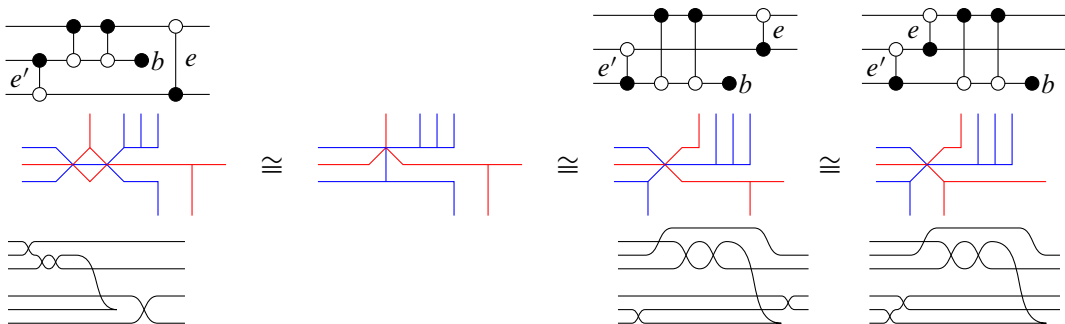
Figure 52: Example of a reflection move near a black lollipop.

obstructing lollipop. In Figure 52, the lollipop b is preventing the vertical edge e from moving to the left in the left picture; after the reflection move, we can now move e through the lollipop b; even better, the edge e is now contained within a subgraph that is a plabic fence, for which we know a recursive procedure to construct the cluster DT transformation [Shen and Weng 2021].

In the front projection, the reflection move is a Legendrian RII move that pulls out a cusp, which is a Legendrian isotopy. Since the reflection move is a weave equivalence, the quiver does not change under such a move. Note that, if e' is the only the vertical edge present in each of the bicolor graphs in Figure 52, then this reflection move recovers the right reflection move in [Shen and Weng 2021]. (A similar picture can be drawn for left reflection moves.)

By realizing the reflection moves as Legendrian RII moves on the front projection, we can define a Legendrian isotopy on Legendrian links associated with shuffle graphs. By construction, between the top region and the bottom region of the initial weave associated with a shuffle graph, the outgoing weave lines inside one of them is always just a half twist. In the front projection, this region can be untangled into a collection of parallel horizontal lines. Thus, we can clockwise rotate every crossing in the other region one by one to this region using just reflection moves (and homotopy) on the front projection. We call this Legendrian isotopy the *half Kálmán loop* $K^{1/2}$. It is not a Legendrian loop and $K^{1/2}$ does not automatically give rise to an automorphism on $\mathcal{M}_1(\Lambda)$: it only gives rise to an isomorphism $K^{1/2} \colon \mathcal{M}_1(\Lambda) \to \mathcal{M}_1(\Lambda')$, where Λ' is the image of Λ under the Legendrian isotopy $K^{1/2}$. In order to make this into an automorphism, we need the involution t induced from the strict contactomorphism $t \colon (x, y, z) \mapsto (-x, y, -z)$ on $\mathbb{R}^3$. By Proposition 4.1, we see that this strict contactomorphism reverses all maps in the quiver representation, which implies that we need to dualize all vector spaces and take transpositions of all the maps. Note that this coincides with the definition of the transposition map $t \colon \mathcal{M}_1(\Lambda') \to \mathcal{M}_1(\Lambda)$ in [Shen and Weng 2021]. All parallel transportation maps are now dualized as well, but the microlocal monodromies and microlocal merodromies remain unchanged and therefore t preserves the cluster structure and is a cluster isomorphism. We define $\mathrm{DT} := t \circ K^{1/2} = K^{1/2} \circ t$ as our candidate for the cluster Donaldson–Thomas transformation for shuffle graphs.

5.3 Edge migration in a plabic fence

Besides the reflection moves, we also need to move vertical edges through regions that locally look like plabic fences, and cluster mutations are needed for this process. In this subsection, we will discuss these moves and prove some basic results about the color change of quiver vertices (green vs red). We begin with a quick review of the meaning of vertex colors, green and red, in a quiver. Fix an initial quiver Q with no frozen vertices. We construct a framed quiver $\widetilde{Q}$ from Q by adding a frozen vertex i' for every vertex i of Q, together with a single arrow pointing from i to i'. Note that, by construction, the exchange matrix of $\widetilde{Q}$ is

$$\widetilde{\epsilon} = \begin{pmatrix} \epsilon & \mathrm{id} \\ -\mathrm{id} & 0 \end{pmatrix},$$

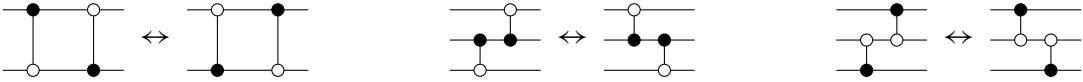where ϵ is the exchange matrix of Q.

Figure 53: Left: the square move in a plabic fence. Center and right: sliding vertical edges of opposite patterns on different levels through each other.

For any mutation sequence μ_i on the quiver Q, we can apply the same mutation sequence μ_i to $\widetilde{Q}$ and get a new quiver $\widetilde{Q}' := \mu_i(Q)$. Note that the unfrozen part of $\widetilde{Q}'$ is identical to $Q' := \mu_i(Q)$. A remarkable property of $\widetilde{Q}'$ is that, for any unfrozen vertex i, $\widetilde{\epsilon}'_{ij'}$ is either nonnegative or nonpositive for all framing frozen vertices j'; this is known as the *sign coherence* phenomenon of c–vectors in cluster theory [Derksen et al. 2010; Gross et al. 2018]. We say a vertex i in Q' is *green* if $\widetilde{\epsilon}'_{ij'}$ is nonnegative for all framing frozen vertices j' and a vertex i in Q' is *red* if $\widetilde{\epsilon}'_{ij'}$ is nonpositive for all framing frozen vertices j'. For a given initial quiver Q (all of whose vertices are green), if a mutation sequence μ_i turns every quiver vertex red, then we say μ_i is a *reddening sequence*. If additionally a reddening sequence μ_i only mutates at green vertices, then we say that μ_i is a *maximal green sequence*. For any fixed initial seed, the cluster Donaldson–Thomas transformation can be captured combinatorially by a reddening sequence [Keller 2017; Goncharov and Shen 2018]. Thus, it is important to keep track of color change of quiver vertices as we perform cluster mutations.

Let us now consider a plabic fence $\mathbb{G}$. By construction, the initial quiver $Q = Q(\mathbb{G})$ is a planar quiver with one unfrozen vertex for each face in $\mathbb{G}$, and the arrows in Q are drawn in a way such that they form a clockwise cycle around a neighboring group of white vertices and form a counterclockwise cycle around a neighboring group of black vertices. If we have two adjacent vertical edges of opposite patterns on the same level, we can exchange them by doing a mutation at the quiver vertex corresponding to the face they bound: this is just the square move. If we have two adjacent vertical edges of opposite patterns not on the same level, then we can slide them through each other without doing any mutation on the quiver. See Figure 53.

On a Legendrian weave, the sliding of edges corresponds to a weave equivalence, whereas the square move can be described by a weave mutation along a long I–cycle, which can be locally described by the movie in Figure 54.

A maximal green sequence on $Q(\mathbb{G})$ can be constructed recursively as follows:

- Take the rightmost vertical edge e of the ⬤◯ pattern and change it to the opposite pattern.
- Move this newly changed vertical edge e to the left, passing all remaining ◯⬤ edges.
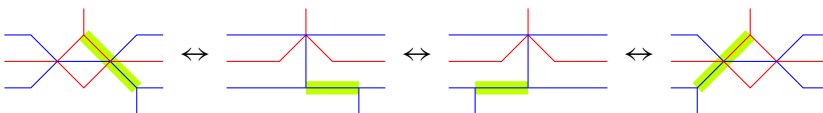
Figure 54: Square move in terms of Legendrian weaves: the first and last moves are weave equivalences; the middle move is a weave mutation.

This iterative process terminates when we run out of vertical edges with a black vertex on top. Note that all mutations in this maximal green sequence come from square moves; this will not be the case for general shuffle graphs. Lastly, here is a result that we will need in the next subsection:

Lemma 5.5 [Shen and Weng 2021, Proposition 4.6] *Each square move turns the mutating vertex from green to red and turns the vertex directly to the right of the mutating vertex (if such a vertex exists) from red back to green. As a result, at the end of each iteration of moving a vertical edge e to the left, the leftmost quiver vertex on the level of e turns red, while the color of every other vertex remain the same.*

5.4 DT transformations for shuffle graphs

Let us construct the cluster DT transformations for shuffle graphs. By Condition (1) of Definition 2.8, if a vertical edge can be placed between the ith and jth horizontal lines with $|i - j| > 1$, then there must be disjoint two continuous regions we can place vertical edges between them, with one on the left and the other one on the right. Let us call them the *left region* and the *right region*, respectively. If $|i - j| = 1$, then there is only one continuous region where we can place vertical edges between the two horizontal lines. The main strategy is to go through all vertical edges of $\mathbb{G}$ one by one from right to left. For each vertical edge e we do one of the following, depending on its location in $\mathbb{G}$:

(I) If e lies on level $(i, i + 1)$:

 (I.1) Apply a reflection move to change the pattern of e to ⦶.

 (I.2) Move e to the left through all ⦶ edges incident to the ith or $(i+1)$st horizontal lines. Note that a cluster mutation occurs whenever we exchange e and a ⦶ edge at the same horizontal level.

(II) If e lies in the right region of level (i, j) with $j - i > 1$:

(II.1) Apply a reflection move to change the pattern of e to ⦶.

(II.2) Move e all the way into the left region between the ith and the jth horizontal lines, and through all ⦶ edges incident to the ith or jth horizontal lines.

Remark 5.6 There is the following subtlety in step (II.2): Since $j - i > 1$, when moving e to the left, we will encounter $j - i - 1$ black lollipops. Each time we encounter a black lollipop, we will try to apply a move similar to Figure 52. If such a move can be applied, then we will get another vertical edge of the same pattern as e, and we need to move this newly changed edge along with e as a group to the left. Moreover, this newly changed edge itself may encounter a black lollipop, too, and consequently introduce another vertical edge into the moving group. This moving group eventually will reach the other side, and we need a way to recover e back as a vertical edge on level (i, j). This can be done by a move mirror to that of Figure 52; see Figure 55.

Note that all moves in Figure 55 correspond to weave equivalences and hence no cluster mutations occur. After the vertical edge e is recovered, we can send the auxiliary vertical edge e' back to where it was
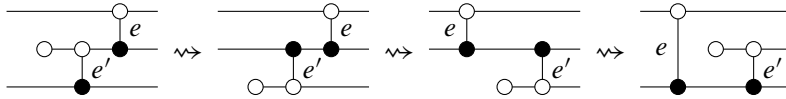
Figure 55: Two reflection moves to cover the vertical edge e.

before, and then do another reflection move to restore the pattern of e'. Note that, upon restoring the location and the pattern of e', the nonplanarity caused by the earlier reflection move (Figure 52, right) will be canceled, and we return to a grid plabic graph after the iteration.

Remark 5.7 There is also a possibility that, although there is a black lollipop b in the way, no vertical edge e' is found and hence it is not possible to perform the move in Figure 52. We claim that in this case we can directly move e through the obstructing horizontal line directly without the need of any weave (cluster) mutations. This follows from the fact that if no vertical edge e' is present, then the incoming weave line corresponding to the gap where e' should have been does not need to be tangled in the weave, and hence we can perform a weave equivalence to move the vertical edge e through. See Figure 56.

(III)   If e lies in the left region between the ith and the jth horizontal lines with $|i - j| > 1$:

(III.1)   Move e all the way to the right region between the ith and the jth horizontal lines so that it becomes the rightmost vertical edge in the plabic graph.

(III.2)   Apply a reflection move to change the pattern of e to $\overset{\circ}{\underset{\bullet}{|}}$.

Note that, in this case, we need to move the vertical edge e to the right before changing its pattern. But, since we are going through vertical edges in $\mathbb{G}$ one by one from right to left, by the time we get to e, all vertical edges to its right must be of the $\overset{\circ}{\underset{\bullet}{|}}$ pattern already. Thus, moving e, which is of the $\overset{\bullet}{\underset{\circ}{|}}$ pattern, to the right through $\overset{\circ}{\underset{\bullet}{|}}$ edges, is completely mirror to step (II.1) before. Figure 57 depicts the reflection moves we need to perform during this process.

We can now conclude the following result:

Theorem 5.8 *Let $\mathbb{G}$ be a shuffle graph. Then* $\text{DT} = t \circ K^{1/2}$ *is the cluster Donaldson–Thomas transformation on* $\mathcal{M}_1(\Lambda)$.

Proof Since t is a cluster isomorphism, it suffices to prove that $K^{1/2}$ gives rise to a reddening sequence. The vertices of the quiver $Q(\mathbb{G})$ are grouped into regions (Figures 50 and 49); we claim that, after each iterative step (I) and step (II) of moving an edge e on level (i, j), the leftmost green vertex of level (i, j)
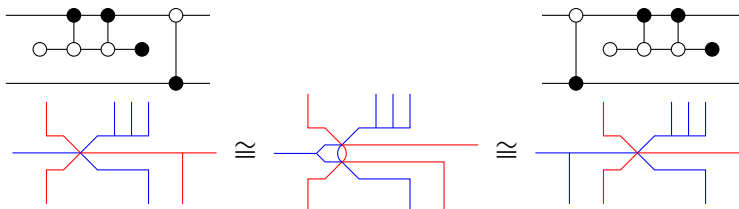

Figure 56: Example of a special case where the edge e' is absent.
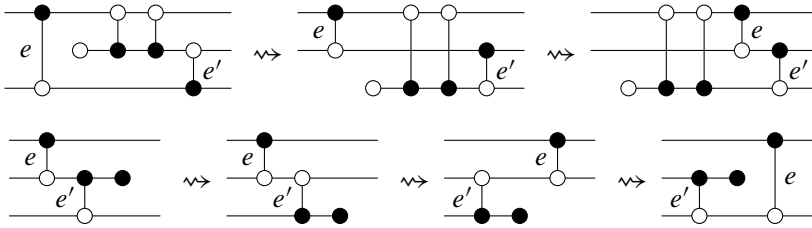
Figure 57: Reflection moves in the (III.1) step: the top row is an example of how to shrink e to a shorter vertical edge, and the bottom row is an example of how to recover e after moving through a horizontal line.

turns red, and after each iterative step (III) of moving an edge e on level (i, j), the rightmost green vertex of level (i, j) turns red. Indeed, the case (I) follows from Lemma 5.5 directly; so it remains to consider (II) and (III).

Let us consider (II) first. In the process of moving a ⦶ edge e on level (i, j) to the left, before we encounter any lollipop, the quiver vertices would change according to Lemma 5.5, turning from green to red when a square move occurs and then turning back to green in the next square move. Let us now consider what happens when we encounter a black lollipop at the kth horizontal line (with $i < k < j$). If we are in the situation of Remark 5.7, then we can directly jump through the whole k^{th} horizontal line without any cluster mutations, and Lemma 5.5 will continue to take care of the rest. It thus remains to consider what happens when we need to do moves according to Remark 5.6. Note that, since the moves in Figure 52 do not induce any cluster mutations, there is no change to the quiver itself. However, the way we branch the quiver is different: the sugar-free hull to the left of edge e was nonrectangular before the moves in Figure 52 but it becomes rectangular after the moves; thus, the corresponding quiver vertex was on the side branch before the moves and relocates itself to the main branch after the moves.

After this quiver vertex is relocated to the main branch (which is a quiver of a plabic fence), we can make use of Lemma 5.5 again. Note that we need to move e as well as the auxiliary edge e' together to the left as a group, and, in that process, there is still a possibility of introducing more edges to that left-moving group. Nevertheless, by induction it is enough to consider what happens when the two-member group e' and e reaches the left white lollipop of the kth horizontal line. By Lemma 5.5, we know that both the quiver vertex v to the right of e and the quiver vertex v' to the right of e' have turned red. Next, under the moves in Figure 55, we restore e to a vertical edge on level (i, j) without any cluster mutations. Finally, we need to send e' back to where it was before, and this process reverses the mutations we did on the level of e': the quiver vertices on that level will turn from green to red and then back to green again one by one;[15] in the end, all quiver vertices on the same level as e' are restored back to green. The iterative step can now continue further to the left on level (i, j), and the color change will again follow Lemma 5.5.

---

[15]Note that we are mutating at red quiver vertices in this process; thus, we are not claiming that the whole mutation sequence is maximal green. On the other hand, such moves are not needed in the case of plabic fences, which is why we can obtain a maximal green sequence.

The case (III) is essentially (II) in reverse. Suppose we are moving a ⬤̥ edge e on level (i, j), and suppose the first white lollipop it encounters is on the kth horizontal line. Let v be the quiver vertex to the right of e, which is the rightmost green on level (i, j) at the beginning of the iterative step. Under the moves in the top row of Figure 57, we move v into level (k, j), and obtain another vertical edge e' of the ⬤̥ pattern. Note that the vertex v' to the right of e' is red at this moment. Now we need to move the vertical edges e and e' to the right: first e', then e. For each square move on level (i, k) (the level of e'), the mutating quiver vertex changes from red to green[16] and stays green afterward. On the other hand, for each square move on level (k, j) (the level of e), the mutating quiver vertex turns from green to red and the one in the next mutation (if it exists) turns from red to green. As a result, when e and e' get to the right end of the kth horizontal line, all quiver vertices on the level of e are red and all quiver vertices on the level of e' are green. After the second row of moves in Figure 57 (which do not involve mutations), we need to send e' back to where it was, and that will make the quiver vertices on the level of e' undergo the reverse sequence of color changes again, restoring all of them back to red. Of course, if there are move vertices further to the right of e, we need to continue moving e rightward, which will make the remaining quiver vertices to the right on level (i, j) turn from red to green and then back to red again one by one. In the end, we see that precisely the quiver vertex v turns red after this iterative step.

In conclusion, since all quiver vertices in $Q(\mathbb{G})$ are either inside the middle region, to the left of a vertical edge in the right region, or to the right of a vertical edge in the left region, we see that, after all vertical edges in $\mathbb{G}$ change from ⬤̥ to ̥⬤, all quiver vertices would turn red. Therefore, $K^{1/2}$ is indeed a reddening sequence. $\qquad\qquad\square$

The statement in Corollary 1.3 about the cluster duality conjecture now follows from [Gross et al. 2018], as our quivers are full-ranked and a DT transformation exists. Finally, we remark that the same argument used in [Gao et al. 2020b] to distinguish infinitely many Lagrangian fillings also works for any shuffle graph whose quiver is mutation equivalent to an acyclic quiver of infinite type. Indeed, the DT transformation will be of infinite order, as will be its square, the Legendrian Kálmán loop.

Example 5.9 Consider the following shuffle graph $\mathbb{G}$, with its quiver depicted and $\Lambda(\mathbb{G})$ at its right:
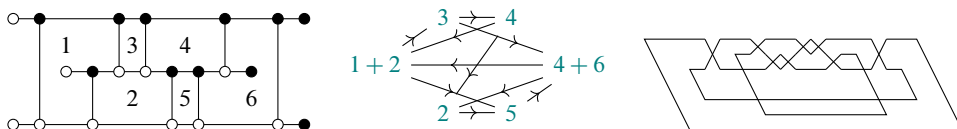
This is mutation equivalent to an acyclic quiver of infinite type, eg consider the mutation sequence $\mu_5 \circ \mu_{4+6} \circ \mu_2 \circ \mu_4 \circ \mu_3 \circ \mu_{4+6}$. Thus, $\Lambda(\mathbb{G})$, which is a max-tb representative in the smooth knot type 10_{161}, admits infinitely many non-Hamiltonian isotopic embedded exact fillings. Note that the

---

[16]These mutations are also not green.

smooth knot type 10_{161} is not a rainbow closure of a positive braid. The reddening sequence realizing DT is

$$(\mu_2 \circ \mu_5 \circ \mu_4 \circ \mu_3 \circ \mu_5 \circ \mu_2) \circ () \circ () \circ (\mu_4) \circ (\mu_5) \circ (\mu_2 \circ \mu_5) \circ (\mu_{4+6} \circ \mu_4)$$
$$\circ (\mu_5 \circ \mu_2 \circ \mu_{1+2} \circ \mu_3 \circ \mu_{4+6} \circ \mu_2 \circ \mu_5),$$

where we have grouped the mutations of each iterative step inside a pair of parentheses, and empty parentheses mean no mutations at that step. The first mutation is μ_5, then μ_2, then μ_{4+6}, and so on until the last three mutations are μ_4, μ_5 and lastly μ_2.

References

[Allegretti 2021]   **D G L Allegretti**, *Stability conditions, cluster varieties, and Riemann–Hilbert problems from surfaces*, Adv. Math. 380 (2021) art. id. 107610   MR Zbl

[Auroux 2007]   **D Auroux**, *Mirror symmetry and T–duality in the complement of an anticanonical divisor*, J. Gökova Geom. Topol. 1 (2007) 51–91   MR Zbl

[Auroux 2009]   **D Auroux**, *Special Lagrangian fibrations, wall-crossing, and mirror symmetry*, from "Geometry, analysis, and algebraic geometry: forty years of the Journal of Differential Geometry" (H-D Cao, S-T Yau, editors), Surv. Differ. Geom. 13, International, Somerville, MA (2009) 1–47   MR Zbl

[Berenstein et al. 2005]   **A Berenstein**, **S Fomin**, **A Zelevinsky**, *Cluster algebras, III: Upper bounds and double Bruhat cells*, Duke Math. J. 126 (2005) 1–52   MR Zbl

[Björner and Brenti 2005]   **A Björner**, **F Brenti**, *Combinatorics of Coxeter groups*, Graduate Texts in Math. 231, Springer (2005)   MR Zbl

[Boalch 2014a]   **P P Boalch**, *Geometry and braiding of Stokes data; fission and wild character varieties*, Ann. of Math. 179 (2014) 301–365   MR Zbl

[Boalch 2014b]   **P Boalch**, *Poisson varieties from Riemann surfaces*, Indag. Math. 25 (2014) 872–900   MR Zbl

[Casals 2022]   **R Casals**, *Lagrangian skeleta and plane curve singularities*, J. Fixed Point Theory Appl. 24 (2022) art. id. 34   MR Zbl

[Casals and Gao 2022]   **R Casals**, **H Gao**, *Infinitely many Lagrangian fillings*, Ann. of Math. 195 (2022) 207–249   MR Zbl

[Casals and Ng 2022]   **R Casals**, **L Ng**, *Braid loops with infinite monodromy on the Legendrian contact DGA*, J. Topol. 15 (2022) 1927–2016   MR Zbl

[Casals and Zaslow 2022]   **R Casals**, **E Zaslow**, *Legendrian weaves: N–graph calculus, flag moduli and applications*, Geom. Topol. 26 (2022) 3589–3745   MR Zbl

[Casals et al. 2020]   **R Casals**, **E Gorsky**, **M Gorsky**, **J Simental**, *Algebraic weaves and braid varieties*, preprint (2020)   arXiv 2012.06931

[Casals et al. 2021]   **R Casals**, **E Gorsky**, **M Gorsky**, **J Simental**, *Positroid links and braid varieties*, preprint (2021)   arXiv 2105.13948

[Casals et al. 2022]   **R Casals**, **E Gorsky**, **M Gorsky**, **I Le**, **L Shen**, **J Simental**, *Cluster structures on braid varieties*, preprint (2022)   arXiv 2207.11607

[Derksen et al. 2010]   **H Derksen**, **J Weyman**, **A Zelevinsky**, *Quivers with potentials and their representations, II: Applications to cluster algebras*, J. Amer. Math. Soc. 23 (2010) 749–790   MR Zbl

[Ekholm et al. 2016]   **T Ekholm**, **K Honda**, **T Kálmán**, *Legendrian knots and exact Lagrangian cobordisms*, J. Eur. Math. Soc. 18 (2016) 2627–2689   MR Zbl

[Etnyre et al. 2013]   **J B Etnyre**, **L L Ng**, **V Vértesi**, *Legendrian and transverse twist knots*, J. Eur. Math. Soc. 15 (2013) 969–995   MR Zbl

[Fock and Goncharov 2006a]   **V V Fock**, **A B Goncharov**, *Cluster $\mathcal{X}$–varieties, amalgamation, and Poisson–Lie groups*, from "Algebraic geometry and number theory" (V Ginzburg, editor), Progr. Math. 253, Birkhäuser, Boston, MA (2006) 27–68   MR Zbl

[Fock and Goncharov 2006b]   **V Fock**, **A Goncharov**, *Moduli spaces of local systems and higher Teichmüller theory*, Publ. Math. Inst. Hautes Études Sci. 103 (2006) 1–211   MR Zbl

[Fock and Goncharov 2009]   **V V Fock**, **A B Goncharov**, *Cluster ensembles, quantization and the dilogarithm*, Ann. Sci. Éc. Norm. Supér. 42 (2009) 865–930   MR Zbl

[Fomin and Zelevinsky 1999]   **S Fomin**, **A Zelevinsky**, *Double Bruhat cells and total positivity*, J. Amer. Math. Soc. 12 (1999) 335–380   MR Zbl

[Fomin and Zelevinsky 2002]   **S Fomin**, **A Zelevinsky**, *Cluster algebras, I: Foundations*, J. Amer. Math. Soc. 15 (2002) 497–529   MR Zbl

[Fomin and Zelevinsky 2003]   **S Fomin**, **A Zelevinsky**, *Cluster algebras, II: Finite type classification*, Invent. Math. 154 (2003) 63–121   MR Zbl

[Fomin et al. 2008]   **S Fomin**, **M Shapiro**, **D Thurston**, *Cluster algebras and triangulated surfaces, I: Cluster complexes*, Acta Math. 201 (2008) 83–146   MR Zbl

[Fomin et al. 2022]   **S Fomin**, **P Pylyavskyy**, **E Shustin**, **D Thurston**, *Morsifications and mutations*, J. Lond. Math. Soc. 105 (2022) 2478–2554   MR Zbl

[Gaiotto et al. 2010]   **D Gaiotto**, **G W Moore**, **A Neitzke**, *Four-dimensional wall-crossing via three-dimensional field theory*, Comm. Math. Phys. 299 (2010) 163–224   MR Zbl

[Gaiotto et al. 2013]   **D Gaiotto**, **G W Moore**, **A Neitzke**, *Spectral networks*, Ann. Henri Poincaré 14 (2013) 1643–1731   MR Zbl

[Galashin and Lam 2023]   **P Galashin**, **T Lam**, *Positroid varieties and cluster algebras*, Ann. Sci. Éc. Norm. Supér. 56 (2023) 859–884   MR Zbl

[Gao et al. 2020a]   **H Gao**, **L Shen**, **D Weng**, *Augmentations, fillings, and clusters*, preprint (2020)   arXiv 2008.10793

[Gao et al. 2020b]   **H Gao**, **L Shen**, **D Weng**, *Positive braid links with infinitely many fillings*, preprint (2020)   arXiv 2009.00499

[Gekhtman et al. 2005]   **M Gekhtman**, **M Shapiro**, **A Vainshtein**, *Cluster algebras and Weil–Petersson forms*, Duke Math. J. 127 (2005) 291–311   MR Zbl

[Gekhtman et al. 2010]   **M Gekhtman**, **M Shapiro**, **A Vainshtein**, *Cluster algebras and Poisson geometry*, Mathematical Surveys and Monographs 167, Amer. Math. Soc., Providence, RI (2010)   MR Zbl

[Goncharov and Kenyon 2013]   **A B Goncharov**, **R Kenyon**, *Dimers and cluster integrable systems*, Ann. Sci. Éc. Norm. Supér. 46 (2013) 747–813   MR Zbl

[Goncharov and Kontsevich 2021]   **A Goncharov**, **M Kontsevich**, *Spectral description of non-commutative local systems on surfaces and non-commutative cluster varieties*, preprint (2021) arXiv 2108.04168

[Goncharov and Shen 2018]   **A Goncharov**, **L Shen**, *Donaldson–Thomas transformations of moduli spaces of G-local systems*, Adv. Math. 327 (2018) 225–348 MR Zbl

[Gross et al. 2015]   **M Gross**, **P Hacking**, **S Keel**, *Birational geometry of cluster algebras*, Algebr. Geom. 2 (2015) 137–175 MR Zbl

[Gross et al. 2018]   **M Gross**, **P Hacking**, **S Keel**, **M Kontsevich**, *Canonical bases for cluster algebras*, J. Amer. Math. Soc. 31 (2018) 497–608 MR Zbl

[Guillermou 2023]   **S Guillermou**, *Sheaves and symplectic geometry of cotangent bundles*, Astérisque 440, Soc. Math. France, Paris (2023) MR Zbl

[Guillermou and Schapira 2014]   **S Guillermou**, **P Schapira**, *Microlocal theory of sheaves and Tamarkin's non displaceability theorem*, from "Homological mirror symmetry and tropical geometry" (R Castano-Bernard, F Catanese, M Kontsevich, T Pantev, Y Soibelman, I Zharkov, editors), Lect. Notes Unione Mat. Ital. 15, Springer (2014) 43–85 MR Zbl

[Guillermou et al. 2012]   **S Guillermou**, **M Kashiwara**, **P Schapira**, *Sheaf quantization of Hamiltonian isotopies and applications to nondisplaceability problems*, Duke Math. J. 161 (2012) 201–245 MR Zbl

[Hacking and Keel 2018]   **P Hacking**, **S Keel**, *Mirror symmetry and cluster algebras*, from "Proceedings of the International Congress of Mathematicians, II: Invited lectures" (B Sirakov, P N de Souza, M Viana, editors), World Sci., Hackensack, NJ (2018) 671–697 MR Zbl

[Hartshorne 1977]   **R Hartshorne**, *Algebraic geometry*, Graduate Texts in Math. 52, Springer (1977) MR Zbl

[Henry and Rutherford 2015]   **M B Henry**, **D Rutherford**, *Ruling polynomials and augmentations over finite fields*, J. Topol. 8 (2015) 1–37 MR Zbl

[Iwaki and Nakanishi 2014]   **K Iwaki**, **T Nakanishi**, *Exact WKB analysis and cluster algebras*, J. Phys. A 47 (2014) art. id. 474009 MR Zbl

[Iwaki and Nakanishi 2016]   **K Iwaki**, **T Nakanishi**, *Exact WKB analysis and cluster algebras, II: Simple poles, orbifold points, and generalized cluster algebras*, Int. Math. Res. Not. 2016 (2016) 4375–4417 MR Zbl

[Jin and Treumann 2017]   **X Jin**, **D Treumann**, *Brane structures in microlocal sheaf theory*, preprint (2017) arXiv 1704.04291

[Kashiwara and Schapira 1985]   **M Kashiwara**, **P Schapira**, *Microlocal study of sheaves*, Astérisque 128, Soc. Math. France, Paris (1985) MR Zbl Correction in 130 (1985) 209

[Kashiwara and Schapira 1990]   **M Kashiwara**, **P Schapira**, *Sheaves on manifolds*, Grundl. Math. Wissen. 292, Springer (1990) MR Zbl

[Keller 2017]   **B Keller**, *Quiver mutation and combinatorial DT-invariants*, preprint (2017) arXiv 1709.03143

[Kontsevich and Soibelman 2010]   **M Kontsevich**, **Y Soibelman**, *Motivic Donaldson–Thomas invariants: summary of results*, from "Mirror symmetry and tropical geometry" (R Castaño Bernard, Y Soibelman, I Zharkov, editors), Contemp. Math. 527, Amer. Math. Soc., Providence, RI (2010) 55–89 MR Zbl

[Lam and Speyer 2022]   **T Lam**, **D E Speyer**, *Cohomology of cluster varieties, I: Locally acyclic case*, Algebra Number Theory 16 (2022) 179–230 MR Zbl

[Muller 2013]   **G Muller**, *Locally acyclic cluster algebras*, Adv. Math. 233 (2013) 207–247 MR Zbl

[Muller 2014]   **G Muller**, *𝒜 = 𝒰 for locally acyclic cluster algebras*, Symmetry Integrability Geom. Methods Appl. 10 (2014) art. id. 094   MR Zbl

[Nadler 2016]   **D Nadler**, *Wrapped microlocal sheaves on pairs of pants*, preprint (2016) arXiv 1604.00114

[Neitzke 2014]   **A Neitzke**, *Cluster-like coordinates in supersymmetric quantum field theory*, Proc. Natl. Acad. Sci. USA 111 (2014) 9717–9724   MR Zbl

[Ng et al. 2017]   **L Ng**, **D Rutherford**, **V Shende**, **S Sivek**, *The cardinality of the augmentation category of a Legendrian link*, Math. Res. Lett. 24 (2017) 1845–1874   MR Zbl

[Ng et al. 2020]   **L Ng**, **D Rutherford**, **V Shende**, **S Sivek**, **E Zaslow**, *Augmentations are sheaves*, Geom. Topol. 24 (2020) 2149–2286   MR Zbl

[Pascaleff and Tonkonog 2020]   **J Pascaleff**, **D Tonkonog**, *The wall-crossing formula and Lagrangian mutations*, Adv. Math. 361 (2020) art. id. 106850   MR Zbl

[Polterovich 1991]   **L Polterovich**, *The surgery of Lagrange submanifolds*, Geom. Funct. Anal. 1 (1991) 198–210   MR Zbl

[Postnikov 2006]   **A Postnikov**, *Total positivity, grassmannians, and networks*, preprint (2006) arXiv math/0609764

[Schnürer 2018]   **O M Schnürer**, *Six operations on dg enhancements of derived categories of sheaves*, Selecta Math. 24 (2018) 1805–1911   MR Zbl

[Serhiyenko et al. 2019]   **K Serhiyenko**, **M Sherman-Bennett**, **L Williams**, *Cluster structures in Schubert varieties in the Grassmannian*, Proc. Lond. Math. Soc. 119 (2019) 1694–1744   MR Zbl

[Shen and Weng 2020]   **L Shen**, **D Weng**, *Cyclic sieving and cluster duality of Grassmannian*, Symmetry Integrability Geom. Methods Appl. 16 (2020) art. id. 067   MR Zbl

[Shen and Weng 2021]   **L Shen**, **D Weng**, *Cluster structures on double Bott–Samelson cells*, Forum Math. Sigma 9 (2021) art. id. e66   MR Zbl

[Shende et al. 2016]   **V Shende**, **D Treumann**, **H Williams**, *On the combinatorics of exact Lagrangian surfaces*, preprint (2016) arXiv 1603.07449

[Shende et al. 2017]   **V Shende**, **D Treumann**, **E Zaslow**, *Legendrian knots and constructible sheaves*, Invent. Math. 207 (2017) 1031–1133   MR Zbl

[Shende et al. 2019]   **V Shende**, **D Treumann**, **H Williams**, **E Zaslow**, *Cluster varieties from Legendrian knots*, Duke Math. J. 168 (2019) 2801–2871   MR Zbl

[Toën and Vaquié 2007]   **B Toën**, **M Vaquié**, *Moduli of objects in dg-categories*, Ann. Sci. École Norm. Sup. 40 (2007) 387–444   MR Zbl

[Yau 2017]   **M-L Yau**, *Surgery and isotopy of Lagrangian surfaces*, from "Proceedings of the sixth international congress of Chinese mathematicians, II" (C-S Lin, L Yang, S-T Yau, J Yu, editors), Adv. Lect. Math. 37, International, Somerville, MA (2017) 143–162   MR Zbl

Department of Mathematics, University of California Davis
Davis, CA, United States
Department of Mathematics, University of California Davis
Davis, CA, United States

casals@math.ucdavis.edu, dweng@ucdavis.edu

Correction to the article
Bimodules in bordered Heegaard Floer homology

ROBERT LIPSHITZ
PETER OZSVÁTH
DYLAN P THURSTON

We correct some errors in our earlier paper (Geom. Topol. 19 (2015) 525–724).

Grading refinement data In the first sentence of the proof of [2, Proposition 3.7] (this is Proposition 3.10 in the arXiv version), the definition of the $(G(\mathcal{Z}), G(\mathcal{Z}))$–set T is wrong and, in particular, the elements $\psi(s) \cdot \psi'(s)^{-1}$ do not lie in T.

To correct this, for each $i = 0, \ldots, 2k$, fix an idempotent $s_i \subset [2k]$ with $|s_i| = i$ and let T_i be the orbit of $\psi(s_i) \cdot \psi'(s_i)^{-1} \in G'(\mathcal{Z})$ under the left action of $G(\mathcal{Z})$, ie

$$T_i = G(\mathcal{Z}) \cdot (\psi(s_i)\psi'(s_i)^{-1}).$$

We claim that T_i is closed under the right action of $G(\mathcal{Z})$ and that the elements $\psi(s) \cdot \psi'(s)^{-1}$ (for $s \subset [2k]$ with $|s| = i$) all lie in T_i.

For the first claim, given $g \in G$ we have

$$M_*\big((\psi(s)\psi'(s)^{-1})g(\psi(s)\psi'(s)^{-1})^{-1}\big) = 0,$$

so $\big((\psi(s)\psi'(s)^{-1})g(\psi(s)\psi'(s)^{-1})^{-1}\big) \in G(\mathcal{Z})$ and hence $(\psi(s)\psi'(s)^{-1})g \in G(\mathcal{Z}) \cdot (\psi(s)\psi'(s)^{-1})$.

The second claim follows from the fact that

$$M_*\big(\psi(s)\psi'(s)(\psi(s_i)\psi'(s_i))^{-1}\big) = 0,$$

so $\psi(s)\psi'(s)(\psi(s_i)\psi'(s_i))^{-1} \in G$.

These claims imply that grading a generator $I(s) \in {}^{\mathcal{A}}[\mathbb{I}]_{\mathcal{A}}$ by $\psi(s) \cdot \psi'(s)^{-1}$ defines a grading on the summand of ${}^{\mathcal{A}}[\mathbb{I}]_{\mathcal{A}}$ with strands grading i by T_i. The rest of the proof of the proposition then goes through unchanged, except working one strands grading i at a time and using T_i in place of T.

The mapping class group action on the category of graded modules Theorem 15 asserts that the bimodules $\widehat{CFDA}(\phi)$ induce an action of the mapping class group on $H_*(\mathrm{Mod}_{\mathcal{A}(\mathcal{Z})})$, which is true, and that this action preserves the subcategories $\mathrm{H}(\widetilde{\mathrm{Mod}_{\mathcal{A}(\mathcal{Z})}})$, which is false. There are two problems with the

second statement:

(1) The grading sets for the bimodules $\widehat{CFDA}(\phi)$ associated to mapping classes are graded by $G(\mathcal{Z})$–sets S_ϕ, so that S_ϕ is isomorphic to $G(\mathcal{Z})$ as a left $G(\mathcal{Z})$–set and as a right $G(\mathcal{Z})$–set, but not as a set with an action of $G(\mathcal{Z}) \times G(\mathcal{Z})^{\mathrm{op}}$. This point is studied further in our later paper [1, Section 5]. Probably this issue could be handled by restricting to the Torelli subgroup, or perhaps a further subgroup.

(2) Even when S_ϕ is isomorphic to $G(\mathcal{Z})$ as a $G(\mathcal{Z}) \times G(\mathcal{Z})^{\mathrm{op}}$–set, the isomorphism is not canonical. So, if M is a module graded by $G(\mathcal{Z})$, then $M \boxtimes_{\mathcal{A}(\mathcal{Z})} \widehat{CFDA}(\phi)$ does not have a canonical grading by $G(\mathcal{Z})$, but rather by a $G(\mathcal{Z})$–set isomorphic to $G(\mathcal{Z})$. This is equivalent to a *relative* grading by $G(\mathcal{Z})$, not an absolute grading by $G(\mathcal{Z})$.

Perhaps it is possible to define canonical absolute gradings on the modules $\widehat{CFDA}(\phi)$ by $G(\mathcal{Z})$ for ϕ in the Torelli group or, perhaps, a nontrivial subgroup of it. Alternatively, one could look for an action of a central extension of the Torelli group. Investigating this would be an interesting future project.

To summarize, the correct statement is the following:

Theorem 15′ *The bimodules $\widehat{CFDA}(\phi)$ induce a weak action of the genus-k mapping class groupoid* $\mathrm{MCG}_0(k)$ *on* $\{H_*(\mathrm{Mod}_{\mathcal{A}(\mathcal{Z})}) \mid \mathrm{genus}(F(\mathcal{Z})) = k\}$.

The proof of Theorem 15′ is identical to the proof of Theorem 15, except that the last paragraph (which was incorrect) is no longer needed.

We thank Andy Manion and Raphael Rouquier for pointing out these mistakes. We also thank the referee for further comments.

Other corrections In the first paragraph of [2, Section 10.1], the notation for the algebras is confused. The algebras $\mathcal{A}$ and $\mathcal{B}$ are, respectively,

$$\mathcal{A} = \left(\iota_0 \underset{\sigma_3}{\overset{\sigma_1}{\underset{\sigma_2}{\rightleftarrows}}} \iota_1 \right) \Big/ \left(\begin{matrix} \sigma_3 \sigma_2, \\ \sigma_2 \sigma_1 \end{matrix} \right), \quad \mathcal{B} = \left(j_0 \underset{\rho_3}{\overset{\rho_1}{\underset{\rho_2}{\rightleftarrows}}} j_1 \right) \Big/ \left(\begin{matrix} \rho_3 \rho_2, \\ \rho_2 \rho_1 \end{matrix} \right)$$

with our usual convention that $\rho_1 \rho_2$ is read left to right, ie this means the arrow ρ_1 followed by the arrow ρ_2. The element σ_{12} is shorthand for $\sigma_1 \sigma_2$ and so on.

We thank Jesse Cohen for pointing out this mistake.

References

[1] **R Lipshitz**, **P S Ozsváth**, **D P Thurston**, *Computing $\widehat{HF}$ by factoring mapping classes*, Geom. Topol. 18 (2014) 2547–2681 MR Zbl

[2] **R Lipshitz**, **P S Ozsváth**, **D P Thurston**, *Bimodules in bordered Heegaard Floer homology*, Geom. Topol. 19 (2015) 525–724 MR Zbl

Department of Mathematics, University of Oregon
Eugene, OR, United States

Department of Mathematics, Princeton University
Princeton, NJ, United States

Department of Mathematics, Indiana University
Bloomington, IN, United States

lipshitz@uoregon.edu,   petero@math.princeton.edu,   dpthurst@indiana.edu

Guidelines for Authors

Submitting a paper to Geometry & Topology

Papers must be submitted using the upload page at the GT website. You will need to choose a suitable editor from the list of editors' interests and to supply MSC codes.

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

Preparing your article for Geometry & Topology

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in LaTeX, preferably using the journal's class file. More information on preparing articles in LaTeX for publication in GT is available on the GT website.

`arXiv` papers

If your paper has previously been deposited on the `arXiv`, we will need its `arXiv` number at acceptance time. This allows us to deposit the DOI of the published version on the paper's `arXiv` page.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Fuzzy or sloppily drawn figures will not be accepted. For labeling figure elements consider the pinlabel LaTeX package, but other methods are fine if the result is editable. If you're not sure whether your figures are acceptable, check with production by sending an email to graphics@msp.org.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

GEOMETRY & TOPOLOGY