Geometry & Topology

Volume 27 (2023)

Issue 8 (pages 2937–3385)

msp

GEOMETRY & TOPOLOGY

msp.org/gt

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

The subscription price for 2023 is US \$740/year for the electronic version, and \$1030/year (+\$70, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP. Geometry & Topology is indexed by Mathematical Reviews, Zentralblatt MATH, Current Mathematical Publications and the Science Citation Index.

Geometry & Topology (ISSN 1465-3060 printed, 1364-0380 electronic) is published 9 times per year and continuously online, by Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840. Periodical rate postage paid at Oakland, CA 94615-9651, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840.

GT peer review and production are managed by EditFLOW® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

Geometry & Topology **27**:8 (2023) 2937–3060
DOI: 10.2140/gt.2023.27.2937
Published: 9 November 2023

Formal groups and quantum cohomology

PAUL SEIDEL

We use chain-level genus-zero Gromov–Witten theory to associate to any closed monotone symplectic manifold a formal group (loosely interpreted), whose Lie algebra is the odd-degree cohomology of the manifold (with vanishing bracket). When taken with coefficients in $\mathbb{F}_p$ for some prime p, the p^{th} power map of the formal group is related to quantum Steenrod operations. The motivation for this construction comes from derived Picard groups of Fukaya categories, and from arithmetic aspects of mirror symmetry.

53D37, 53D45; 14L05

1 Introduction

This paper is concerned with aspects of genus-zero Gromov–Witten theory, which are specifically of interest if one works with integer or mod p cohomological coefficients. There is a shared context between this and arithmetic aspects of Fukaya categories — see for instance Alston and Amorim [2], Evans and Lekili [18], Lekili and Perutz [41] and Lekili and Polishchuk [42] — even though we do not work on a categorical level. Instead, our constructions will resemble those of certain chain-level structures, and of cohomology operations, in algebraic topology.

1a Background

Gromov–Witten theory on a closed symplectic manifold X can be axiomatized as a cohomological field theory (see Kontsevich and Manin [38]), which means that operations are parametrized by Deligne–Mumford moduli spaces of curves. We will only consider genus-zero curves, where the notion of cohomological field theory is related to ones from classical topology: namely, one can start with the little disc operad (May [51, Chapter IV]), then enlarge it to the framed little disc operad (Getzler [28]), and finally trivialize the circle action (Drummond-Cole [17]) to obtain the genus-zero Deligne–Mumford operad. It is important for this paper to work on the chain level. An example of a chain-level construction is the quantum A_∞–ring structure (Ruan and Tian [61]), which refines the small quantum product. In abstract terms, this comes from mapping Stasheff associahedra to Deligne–Mumford spaces, compatibly with the operad structures.

To define the genus-zero cohomological field theory for a general X, one usually has to work with coefficient rings containing $\mathbb{Q}$, because of the multivalued perturbations involved in making moduli spaces regular. However, in the special case where X is weakly monotone,[1] the relevant Gromov–Witten invariants, which count genus-zero curves with ≥ 3 marked points in a given homology class, can be defined over $\mathbb{Z}$. If one reduces coefficients to a finite field $\mathbb{F}_p$, there are two obvious constructions of cohomology operations. One can use the relation with the little disc operad to obtain analogues of the Cohen operations [14] on the homology of double loop spaces. For ease of reference, let's call these quantum Cohen operations. The second approach is to introduce quantum Steenrod operations, which were proposed in Fukaya [21] and have attracted some recent attention in Wilkins [74]. These are both facets of a

[1] Also called semipositive; see eg McDuff and Salamon [54, Section 6.4].

common story, which involves the equivariant cohomology of Deligne–Mumford space with $p+1$ marked points, with respect to the action of the symmetric group Sym_p permuting all but one point.

We take our bearings from both the "one-dimensional" quantum A_∞–structure and the "two-dimensional" quantum Cohen and Steenrod operations. To thread our way between the two, we use another family of moduli spaces, which come from the convolution theory of Lagrangian correspondences, as in Bottman [6], Bottman and Wehrheim [9], Fukaya [22] and Ma'u, Wehrheim and Woodward [48]. They map to Deligne–Mumford spaces, and on the other hand, their boundary structure is governed by Stasheff associahedra. The effect of using (the simplest of) these spaces is to equip the set of quantum Maurer–Cartan solutions with a multiplicative structure. After reduction mod p, that structure will admit a partial description in terms of a specific quantum (Cohen or) Steenrod operation.

1b Algebraic terminology

Before continuing the discussion, we need to recall some definitions. In a "functor of points" approach, an object is often described as a functor from a class of "coefficient rings" to sets. We use the following coefficient rings, familiar from the theory of formal schemes and from deformation theory.

Definition 1.1 An adic ring is a nonunital commutative ring N such that the map $N \to \varprojlim_m N/N^m$ is an isomorphism. In other words, $\bigcap_m N^m = 0$, and N is complete with respect to the topology given by the decreasing filtration $\{N^m\}$. Note that one can adjoin a unit, forming the augmented ring $\mathbb{Z}1 \oplus N$, which contains N as an ideal.

Example 1.2 Standard examples are $N = q\mathbb{Z}[\![q]\!]$ (power series with zero constant term) or its truncations $N = q\mathbb{Z}[q]/q^{m+1}$. We can also use field coefficients, for instance taking $N = q\mathbb{F}_p[\![q]\!]$, which simplifies the algebraic behavior slightly. An example with "unequal characteristic" is $N = p\mathbb{Z}_p$, the maximal ideal in the ring of p–adic integers, where $N/N^m = \mathbb{Z}/p^{m-1}$.

Definition 1.3 A "formal group" is a functor from adic rings to groups.

This is somewhat weaker than the classical notion of formal group as in Lazard [39]: there, one imposes additional conditions on the functor, leading to representability results in an appropriate category of formal schemes. In our application, we will be truncating what should really be an object of derived geometry, and representability in the classical sense is not expected to hold. For simplicity, we have chosen to ignore the issue, resulting in the definition given above.

As mentioned before, adic rings are a standard way to formulate deformation problems; see Schlessinger [63]. The specific problem relevant for us is the following. Let $\mathcal{A}$ be an A_∞–ring; see Section 2c for our conventions. Given an adic ring N, let $\mathcal{A} \mathbin{\widehat{\otimes}} N$ be the inverse limit of tensor products $\mathcal{A} \otimes (N/N^m)$. We consider solutions $\gamma \in \mathcal{A}^1 \mathbin{\widehat{\otimes}} N$ of the (generalized) Maurer–Cartan equation

$$\sum_{d \geq 1} \mu_{\mathcal{A}}^d(\gamma, \ldots, \gamma) = 0. \tag{1-1}$$

Two such solutions $\gamma, \tilde{\gamma}$ are considered equivalent if there is an $h \in \mathcal{A}^0 \mathbin{\widehat{\otimes}} N$ such that

$$\sum_{p,q} \mu_{\mathcal{A}}^{p+q+1}(\overbrace{\gamma, \ldots, \gamma}^{p}, h, \overbrace{\tilde{\gamma}, \ldots, \tilde{\gamma}}^{q}) = \gamma - \tilde{\gamma}. \tag{1-2}$$

Definition 1.4 $\mathrm{MC}(\mathcal{A}; N)$ is the set of equivalence classes of Maurer–Cartan elements in $\mathcal{A} \mathbin{\widehat{\otimes}} N$. This is functorial in N, giving a functor $\mathrm{MC}(\mathcal{A})$ from adic rings to sets.

If $N^2 = 0$, (1-1) reduces to $\mu_{\mathcal{A}}^1(\gamma) = 0$, and (1-2) to $\mu_{\mathcal{A}}^1(h) = \gamma - \tilde{\gamma}$. Hence, in this case $\mathrm{MC}(\mathcal{A}; N) = H^1(\mathcal{A}; N)$ is the cohomology with coefficients in the abelian group N. Correspondingly, the general $\mathrm{MC}(\mathcal{A}; N)$ can be viewed as nonlinear analogues of cohomology groups. Note that what we are studying is not the deformation theory of $\mathcal{A}$ as an A_∞–ring: instead, it can be viewed as the deformation theory of the free module $\mathcal{A}$, inside the dg category of A_∞–modules.

1c The formal group structure

With this in mind, let's return to symplectic geometry. To keep the formalism in the simple form set up above (avoiding Novikov rings), we will assume that our symplectic manifold X is monotone (rather than weakly monotone), which means that its symplectic form satisfies

$$[\omega_X] = \delta c_1(X) \in H^2(X; \mathbb{R}) \quad \text{for some } \delta > 0. \tag{1-3}$$

Take a suitable chain complex $\mathcal{C} = C^*(X)$ representing its integral cohomology, equipped with the quantum A_∞–structure $\mu_{\mathcal{C}}$. Note that the quantum A_∞–structure is only $\mathbb{Z}/2$–graded; hence, the definition above should be interpreted so that Maurer–Cartan elements are taken in $\mathcal{C}^{\mathrm{odd}} \mathbin{\widehat{\otimes}} N$, and correspondingly, the entire odd-degree cohomology of X appears. Let $\mathrm{MC}(X; N) = \mathrm{MC}(\mathcal{C}; N)$ be the set of equivalence classes of Maurer–Cartan solutions. One can think of this as the deformation theory of the diagonal Δ_X as an object of the Fukaya category $\mathcal{F}(X \times \bar{X})$, where $\bar{X}$ means that we have reversed the sign of the symplectic form. In other words, deformations are "bounding cochains" for Δ_X in the sense of Fukaya, Oh, Ohta and Ono [23; 24]. If the closed–open map is an isomorphism, one can also think of this theory as deformations of the identity functor on $\mathcal{F}(X)$, which describes the formal neighborhood of the identity in the "automorphism group" of that category. From the composition of automorphisms, one would expect additional structure, and indeed:

Proposition 1.5 *The functor* $\mathrm{MC}(X)$ *has the canonical structure of a "formal group".*

As mentioned above, if one makes suitable assumptions on the closed–open map, this structure has an explanation purely within homological algebra. If one drops that assumption, one could still obtain the group structure by looking at $\mathcal{F}(X \times \bar{X})$ together with its monoidal structure (in a suitable sense, which we will not try to make precise) given by convolution of correspondences — see Bottman and Wehrheim [9] and Ma'u, Wehrheim and Woodward [48]; another approach is Fukaya [22] and Lekili and Lipyanskiy [40]. Compared to those constructions, the definition given here (which avoids talking about Fukaya categories or Lagrangian correspondences) is less general but more direct, and hence more amenable to computations.

Proposition 1.6 *The groups* $\mathrm{MC}(X; N)$ *are commutative if* $N^3 = 0$. *They are also commutative if* $N^4 = 0$ *provided that*, *additionally*, $H^*(X; \mathbb{Z})$ *is torsion-free.*

Commutativity mod N^3 is not surprising: it amounts to the well-known fact that the Lie bracket on cohomology, which exists for any algebra over the little disc operad, becomes zero for cohomological field theories. For general algebraic reasons (formal exponentiation), one expects commutativity to hold always if N is an algebra over $\mathbb{Q}$; and the same should be true if N is an algebra over $\mathbb{F}_p$ and $N^p = 0$. In contrast, the origin of the second part of Proposition 1.6 is more geometric: it reflects an explicit (if poorly understood, partly due to a lack of examples) enumerative obstruction to commutativity.

Remark 1.7 Our construction focuses on the odd-degree cohomology of X. One could try to include even-degree classes by enlarging the notion of formal group to its derived counterpart, which in our terms means allowing N to be a commutative dg (or maybe better simplicial) ring. Another potential use of even-degree classes (with different enumerative content) would be as "bulk insertions" at points in arbitrary position, as in big quantum cohomology. Note however that, for classes of degree > 2, the standard algebraic formalism of "bulk insertions" involves dividing by factorials. Hence, it would have to be modified for our applications. Neither direction will be attempted in this paper.

1d Quantum Steenrod operations

Fix a prime p. The quantum Steenrod operation, in a form slightly simplified by the monotonicity assumption (1-3), is a map

$$Q\,\mathrm{St}_{X,p} = \sum_A Q\,\mathrm{St}_{X,p,A}\colon H^*(X;\mathbb{F}_p) \to H^*(X;\mathbb{F}_p) \otimes H^*_{\mathbb{Z}/p}(\mathbb{F}_p), \tag{1-4}$$

with

$$Q\,\mathrm{St}_{X,p,A}\colon H^l(X;\mathbb{F}_p) \to \big(H^*(X;\mathbb{F}_p) \otimes H^*_{\mathbb{Z}/p}(\mathbb{F}_p)\big)^{pl-2c_1(A)}.$$

Here, $H^*_{\mathbb{Z}/p}(\mathbb{F}_p)$ is the group cohomology of the cyclic group with coefficients mod p, which is one-dimensional in each degree. We fix generators

$$H^*_{\mathbb{Z}/p}(\mathbb{F}_p) = \mathbb{F}_p[t,\theta], \quad \text{with } |t| = 2,\ |\theta| = 1. \tag{1-5}$$

The notation here requires some explanation. For $p = 2$, we have $\theta^2 = t$ (or $\theta = t^{1/2}$), so the two generators are not independent. For $p > 2$, it is implicit that our description is as a graded commutative algebra, so $\theta^2 = 0$. The sum in (1-4) is over $A \in H_2(X;\mathbb{Z})$, and the notation $c_1(A)$ is shorthand for integrating the first Chern class of X over A. The classical Steenrod operations [70] are encoded in the $A = 0$ term. More precisely, if we write $\mathrm{St}_{X,p} = Q\,\mathrm{St}_{X,p,0}$, the relation with the classical notation is that

$$\mathrm{St}_{X,p}(x) = \begin{cases} \displaystyle\sum_i \mathrm{Sq}^i(x) t^{(|x|-i)/2} & \text{if } p = 2, \\ \displaystyle(-1)^* \Big(\frac{p-1}{2}!\Big)^{|x|} \sum_i (-1)^i P^i(x)\, t^{(|x|-2i)(p-1)/2} & \\ \qquad\qquad + \theta(\text{terms involving } \beta P^i) & \text{if } p > 2, \end{cases} \tag{1-6}$$

where β is the Bockstein, and

$$* = \frac{|x|(|x|-1)}{2}\frac{p-1}{2}. \tag{1-7}$$

When handling the constants in (1-6) in practice, one should bear in mind that

$$\Big(\frac{p-1}{2}!\Big)^2 \equiv (-1)^{(p+1)/2} \bmod p. \tag{1-8}$$

See [70, Lemma 6.3]. For instance, if $|x|$ is even and $p > 2$,

$$\begin{aligned}(1\text{-}9)\quad t^0 \text{ term of } \mathrm{St}_{X,p}(x) &= (-1)^* \Big(\frac{p-1}{2}!\Big)^{|x|} (-1)^{|x|/2} P^{|x|/2}(x) \\ &= (-1)^{|x|/2\,(p-1)/2} (-1)^{|x|/2\,(p+1)/2} (-1)^{|x|/2} P^{|x|/2}(x) \\ &= P^{|x|/2}(x) = x^p.\end{aligned}$$

Definition 1.8 Define an endomorphism $Q\Xi_{X,p}$ of $H^{\mathrm{odd}}(X;\mathbb{F}_p)$ by

$$(1\text{-}10)\quad Q\Xi_{X,p}(x) = \begin{cases} \text{the } t^{1/2} \text{ (or } \theta\text{) component of } Q\,\mathrm{St}_{X,2}(x) & \text{if } p = 2, \\ \Big(\frac{p-1}{2}!\Big)^{-1} \text{ times the } t^{(p-1)/2}\text{–component of } Q\,\mathrm{St}_{X,p}(x) & \text{if } p > 2. \end{cases}$$

To recapitulate, this has the form

$$(1\text{-}11)\qquad Q\Xi_{X,p} = \sum_A Q\Xi_{X,p,A},$$

with

$$Q\Xi_{X,p,A} \colon H^l(X;\mathbb{F}_p) \to H^{pl-(p-1)-2c_1(A)}(X;\mathbb{F}_p),$$

and where the classical component is

$$(1\text{-}12)\qquad \Xi_{X,p}(x) = Q\Xi_{X,p,0}(x) = \begin{cases} \mathrm{Sq}^{|x|-1}(x) & \text{if } p = 2, \\ P^{(|x|-1)/2}(x) & \text{if } p > 2. \end{cases}$$

1e The p^{th} power maps

Let's return to the formal group $\mathrm{MC}(X)$. The group structure gives rise to m^{th} power (meaning the m–fold product) maps for each $m \geq 1$, which are functorial endomorphisms of $\mathrm{MC}(X;N)$ for any N.

Theorem 1.9 *The power maps of prime order fit into a diagram*

$$(1\text{-}13)\qquad \begin{tikzcd} \mathrm{MC}(X;q\mathbb{F}_p[q]/q^{p+1}) \arrow[rr, "p^{\text{th}} \text{ power of the formal group}"] \arrow[d, "\text{projection}"'] & & \mathrm{MC}(X;q\mathbb{F}_p[q]/q^{p+1}) \\ \mathrm{MC}(X;q\mathbb{F}_p[q]/q^2) \arrow[d, equal] & & \mathrm{MC}(X;q^p\mathbb{F}_p[q]/q^{p+1}) \arrow[u, "\text{inclusion}"'] \arrow[d, equal] \\ H^{\mathrm{odd}}(M;\mathbb{F}_p) \arrow[rr, "Q\Xi_{X,p}"] & & H^{\mathrm{odd}}(M;\mathbb{F}_p) \end{tikzcd}$$

Remark 1.10 Because of the monotonicity of X and the grading of our operations, see (1-11), one always has

$$Q\Xi_{X,p}(x) = \Xi_{X,p}(x) = x \quad \text{for } x \in H^1(X;\mathbb{F}_p). \tag{1-14}$$

For comparison, consider the formal completion $\widehat{\mathbb{G}}_m$ of the multiplicative group. In a local coordinate $1+z \in \widehat{\mathbb{G}}_m$, the p^{th} power map is

$$\overbrace{z \bullet \cdots \bullet z}^{p} = (1+z)^p - 1 = z^p + p(\text{something}) = z \quad \text{for } z \in \mathbb{F}_p, \tag{1-15}$$

which matches what we have seen in (1-14). There is a categorical explanation for the occurrence of the multiplicative group. Recall that $H^1(X;\mathbb{G}_m)$ classifies flat line bundles over X. The definition of the Fukaya category $\mathcal{F}(X)$ includes having the Lagrangian submanifolds equipped with flat bundles. By tensoring with the restriction of flat line bundles on X, one gets an action of $H^1(X;\mathbb{G}_m)$ on the Fukaya category. For us, it is better to think of the action as being given by the trivial Lagrangian correspondence, namely the diagonal in $X \times \bar{X}$, equipped with a flat line bundle. From that viewpoint, one can pass to the formal completion: one has a formal family of objects in $\mathcal{F}(X \times \bar{X})$, which consists of the diagonal together with a formal deformation of the trivial line bundle; that gives rise to a deformation of the identity functor on $\mathcal{F}(X)$; and composition of such deformations corresponds to the tensor product of line bundles. Of course, within the present framework this discussion is of very limited concrete use: the known examples of monotone symplectic manifolds with nontrivial H^1 (obtained by combining Reznikov [60] and Millson [55], see Fine and Panov [19] for a discussion) are somewhat esoteric.

Example 1.11 Let $X \subset \mathbb{C}P^1 \times \mathbb{C}P^3$ be a hypersurface of bidegree $(1,2)$, which has odd cohomology $H^3(X;\mathbb{F}_p) = \mathbb{F}_p^2$ for any p. Then $Q\Xi_{X,2} = \mathrm{id}$, by a computation from [74, Section 8]. More generally, each $Q\Xi_{X,p}$ is a multiple of the identity. Here are the results for the first few primes:

(1-16)

p	2	3	5	7	11	13	17	19	23	29	31	37	41
$Q\Xi_p/\mathrm{id}$	−1	−1	1	0	−4	−2	2	4	0	−2	0	−10	10

The entries lie in $\mathbb{F}_p$, and we have chosen integer representatives with the least absolute value (with some fudging for $p=2$). Those integers are meaningful: they are the q^p coefficients of the modular form [43, Newform 15.2.a.a]

$$\eta(q)\eta(q^3)\eta(q^5)\eta(q^{15}), \quad \text{where } \eta(q) = q^{1/24}\prod_{n=1}^{\infty}(1-q^n). \tag{1-17}$$

One can interpret this observation via mirror symmetry and arithmetic geometry. The (conjectural, but supported by superpotential computations) statement is that a specific elliptic curve appears in the mirror geometry, and hence is encoded in the Fukaya category of X. Correspondingly, the automorphism group of the Fukaya category would contain the derived automorphism group of that curve, and in particular, the product of two copies of the curve itself. What we see in (1-16) is the leading coefficient of the p^{th} power map of the formal group law of the elliptic curve. For general number theory reasons, this is closely related to counting $\mathbb{F}_p$–points on the curve, and the appearance of (1-17) is an instance of the modularity of elliptic curves. For further discussion, see Example 9.11 and Conjecture 9.12.

The computation underlying Example 1.11 turns out to involve only those quantum Steenrod operations which can ultimately (using forthcoming work of Wilkins and the author) be reduced to ordinary Gromov–Witten invariants. To push the understanding of $Q\Xi_{X,p}$ further, one would have to study the contribution of p–fold covered curves, which is beyond our scope here.

Example 1.12 Let $X \subset \mathbb{C}P^1 \times \mathbb{C}P^5$ be a hypersurface of bidegree $(1,2)$. In this case, $Q\Xi_{X,p}$ is unknown. The answer involves stable maps to X with first Chern number $2p-2$. The difficulty is that there are points in the relevant space of stable maps which have $\mathbb{Z}/p$ isotropy groups.

1f Structure of the paper

In order to make the underlying ideas appear clearly, the paper is set up as follows. Most of the time (Sections 2–6) we work in an abstract operadic framework. In principle, one could aim to prove that quantum cohomology is an instance of this general setup, but that would overshoot the desired target somewhat. Instead, we will explain (in Section 7) how to convert the previous arguments into symplectic terms, in a more ad hoc way. In Section 8, we outline an alternative approach to parts of the construction, based on Fukaya [22]. After that, Section 9 is a bit of an outlier: it is concerned with computational techniques for quantum Steenrod operations, and is formulated in a language much closer to standard Gromov–Witten theory. At this point, we should make one apology for the paper. Because of the complexity of the formulae involved, signs are sometimes not worked out, which we signal by $\pm$; however, we have made sure that signs are given at key points. Part of this involves spelling out certain conventions for equivariant cohomology, which is done in Section 10.

Acknowledgements I would like to thank Nate Bottman and Kenji Fukaya for providing important insights into moduli spaces and Fukaya categories; Nicholas Wilkins for many conversations about quantum Steenrod operations; John Pardon, Bjorn Poonen and Andrew Sutherland for teaching me bits of arithmetic geometry; Alessio Corti, Vasily Golyshev and Victor Przyjalkowski for useful information about mirror symmetry and periods; and Mikhail Kapranov for pointing out related homological algebra results.

This research was partially supported by the Simons Foundation, through a Simons Investigator award as well as the Simons Collaboration in Homological Mirror Symmetry; and by the National Science Foundation, through award DMS-1904997. I would also like to thank Columbia University, Princeton University and the Institute for Advanced Study, for generous visiting appointments during which I worked on this paper.

2 Maurer–Cartan theory

After some introductory remarks about solutions of the Maurer–Cartan equations in general A_∞–rings, we turn to a specific situation, namely the induced A_∞–structure on Hochschild cochains. Maurer–Cartan solutions in Hochschild cochains carry a formal group structure, which can be considered as a purely algebraic counterpart of our main construction. This algebraic viewpoint will not really be used later on: we include it here for expository purposes, and also because it would provide the background for linking the results in this paper to the Fukaya category. To make things more intuitive from a classical homological algebra viewpoint, we will take the A_∞–structures to be $\mathbb{Z}$–graded in this section, even though, as mentioned before, the quantum A_∞–structure is only $\mathbb{Z}/2$–graded.

2a A_∞–structures

To clarify our conventions, let's spell out the definition of an A_∞–ring. This is a free graded abelian group $\mathcal{A}$ is with multilinear operations $\{\mu^d_{\mathcal{A}}\}$, $d \geq 1$, which satisfy the A_∞–associativity relations

$$0 = \sum_{ij} (-1)^{\maltese_i} \mu_{\mathcal{A}}^{d-j+1}(a_1, \dots, \mu_{\mathcal{A}}^j(a_{i+1}, \dots, a_{i+j}), \dots, a_d). \tag{2-1}$$

Here, $\maltese_i = \|a_1\| + \dots + \|a_i\|$, where $\|a\| = |a| - 1$ is the reduced degree; both will be standing notation from now on. If we consider $\mathcal{A}$ as a chain complex with differential $d_{\mathcal{A}} = -\mu^1_{\mathcal{A}}$, the associative algebra structure on $H^*(\mathcal{A})$ is induced by the chain-level product

$$a_1 \cdot a_2 = (-1)^{|a_1|}\mu^2_{\mathcal{A}}(a_1, a_2). \tag{2-2}$$

From the overall "A_∞–lingo", the notions of A_∞–homomorphism and homotopy between such homomorphisms will be the ones that occur most frequently in our discussion. Homotopy admits the following useful interpretation. Take the following dg ring (cochains on the interval as a simplicial complex, with the Alexander–Whitney product):

$$\begin{aligned}
&\mathcal{I} = \mathbb{Z}u \oplus \mathbb{Z}\tilde{u} \oplus \mathbb{Z}v, \quad \text{with } |u| = |\tilde{u}| = 0,\ |v| = 1,\\
&u^2 = u, \quad \tilde{u}^2 = \tilde{u}, \quad \tilde{u}u = u\tilde{u} = 0, \quad uv = v = v\tilde{u} \quad (\text{and hence } vu = \tilde{u}v = 0),\\
&d_{\mathcal{I}}u = v, \quad d_{\mathcal{I}}\tilde{u} = -v.
\end{aligned} \tag{2-3}$$

If $\mathcal{A}$ is an A_∞–ring, the tensor product $\mathcal{A} \otimes \mathcal{I}$ inherits the same structure, with

$$\begin{aligned}
&\mu^1_{\mathcal{A}\otimes\mathcal{I}}(a \otimes x) = \mu^1_{\mathcal{A}}(a) \otimes x + (-1)^{|a|} a \otimes d_{\mathcal{I}}x,\\
&\mu^d_{\mathcal{A}\otimes\mathcal{I}}(a_1 \otimes x_1, \dots, a_d \otimes x_d) = (-1)^* \mu^d_{\mathcal{A}}(a_1, \dots, a_d) \otimes x_1 \cdots x_d \quad \text{for } d \geq 2,
\end{aligned} \tag{2-4}$$

where $* = \sum_{i>j} \|a_i\| \cdot |x_j|$. This A_∞–structure is compatible with the projections

$$\begin{tikzcd} \mathcal{A} \otimes \mathcal{I} \arrow[r, shift left, "\text{project to } \mathcal{A} \otimes \mathbb{Z}u"] \arrow[r, shift right, "\text{project to } \mathcal{A} \otimes \mathbb{Z}\tilde{u}"'] & \mathcal{A}. \end{tikzcd} \tag{2-5}$$

Two A_∞–homomorphisms $\tilde{\mathcal{A}} \to \mathcal{A}$ are homotopic if and only if they can be obtained from a common homomorphism $\tilde{\mathcal{A}} \to \mathcal{A} \otimes \mathcal{I}$ by composing with (2-5). We will often use the following fact:

Lemma 2.1 *Let $\mathcal{F}\colon \tilde{\mathcal{A}} \to \mathcal{A}$ be an A_∞–homomorphism such that the linear term $\mathcal{F}^1$ is a chain homotopy equivalence (in view of our freeness assumption, that will be the case whenever it's a quasi-isomorphism). Then $\mathcal{F}$ has an inverse up to homotopy.*

Unitality conditions, while not always strictly necessary, are both convenient for the theory and satisfied in most applications (including ours). A homology unit for $\mathcal{A}$ is a cocycle $e_{\mathcal{A}} \in \mathcal{A}^0$ such that the products

$$a \mapsto a \cdot e_{\mathcal{A}} = (-1)^{|a|}\mu^2_{\mathcal{A}}(a, e_{\mathcal{A}}) \quad \text{and} \quad a \mapsto e_{\mathcal{A}} \cdot a = \mu^2_{\mathcal{A}}(e_{\mathcal{A}}, a) \tag{2-6}$$

are homotopic to the identity (when working over a field, one asks that these products induce the identity on cohomology, but that is obviously inadequate over $\mathbb{Z}$; the notion used here goes back to [44, Definition 7.3]). One says that $e_{\mathcal{A}}$ is a strict unit if: the inclusion $\mathbb{Z}e_{\mathcal{A}} \to \mathcal{A}^0$ splits, as a map of abelian groups; the maps (2-6) are equal to the identity; and in addition, all operations $\mu^d_{\mathcal{A}}(\dots, e_{\mathcal{A}}, \dots)$, $d \geq 3$, are zero. The following is [45, Theorem 3.7 and Remark 3.8]:

Lemma 2.2 *Given any homologically unital A_∞–ring $\mathcal{A}$, there is a strictly unital one $\tilde{\mathcal{A}}$ and an inclusion $\mathcal{A} \hookrightarrow \tilde{\mathcal{A}}$, compatible with the A_∞–structures, which is a chain homotopy equivalence.* (Note that by Lemma 2.1, we then also have an inverse A_∞–functor $\mathcal{F}\colon \tilde{\mathcal{A}} \to \mathcal{A}$, such that $\mathcal{F}^1$ is a chain homotopy equivalence.)

The result in [45] is more explicit: one can enlarge the A_∞–structure to $\tilde{\mathcal{A}} = \mathcal{A} \oplus \mathbb{Z}h \oplus \mathbb{Z}e_{\tilde{\mathcal{A}}}$, where $e_{\tilde{\mathcal{A}}}$ is the strict unit, and

$$\begin{aligned} &\mu^1_{\tilde{\mathcal{A}}}(h) \in e_{\tilde{\mathcal{A}}} + \mathcal{A}^0, \\ &\mu^d_{\tilde{\mathcal{A}}}(\mathcal{A} \oplus \mathbb{Z}h, \dots, \mathcal{A} \oplus \mathbb{Z}h) \subset \mathcal{A} \quad \text{for } d \geq 2. \end{aligned} \tag{2-7}$$

This has a consequence which we find useful to state, even though it goes slightly beyond the limits of our current terminology. Introduce an A_∞–category with two objects Y and $\tilde{Y}$, morphism spaces

$$\begin{aligned} &\hom(Y, Y) = \hom(Y, \tilde{Y}) = \hom(\tilde{Y}, Y) = \mathcal{A}, \\ &\hom(\tilde{Y}, \tilde{Y}) = \tilde{\mathcal{A}}, \end{aligned} \tag{2-8}$$

and with all A_∞–structures inherited from $\tilde{\mathcal{A}}$ (the second part of (2-7) ensures that this makes sense). The two objects are quasi-isomorphic, and so we arrive at the following:

Lemma 2.3 *Given any homologically unital A_∞–ring $\mathcal{A}$, there is a homologically unital A_∞–category with two objects such that*

- *the endomorphism ring of the first object is $\mathcal{A}$,*
- *the endomorphism ring of the second object is strictly unital, and*
- *the two objects are mutually quasi-isomorphic.*

2b Maurer–Cartan elements

We have already mentioned the notions of Maurer–Cartan element (1-1) and of equivalence between such elements (1-2). Given an A_∞–homomorphism $\mathcal{F}\colon \tilde{\mathcal{A}} \to \mathcal{A}$, we define the induced map $\mathrm{MC}(\mathcal{F}; N)\colon \mathrm{MC}(\tilde{\mathcal{A}}; N) \to \mathrm{MC}(\mathcal{A}; N)$ by

$$\tilde{\gamma} \mapsto \gamma = \sum_d \mathcal{F}^d(\tilde{\gamma}, \dots, \tilde{\gamma}). \tag{2-9}$$

The basic results (the second is a consequence of the first and Lemma 2.1) are:

Lemma 2.4 *Homotopic A_∞–homomorphisms induce the same map* $\mathrm{MC}(\widetilde{\mathcal{A}}; N) \to \mathrm{MC}(\mathcal{A}; N)$.

Lemma 2.5 *Suppose that we have an A_∞–homomorphism $\widetilde{\mathcal{A}} \to \mathcal{A}$, whose linear part is a chain homotopy equivalence. Then the induced map* $\mathrm{MC}(\widetilde{\mathcal{A}}; N) \to \mathrm{MC}(\mathcal{A}; N)$ *is bijective.*

One can think of equivalence of Maurer–Cartan elements in several ways. In terms of the A_∞–structure (2-4),

$$\gamma \otimes u + \tilde{\gamma} \otimes \tilde{u} + h \otimes v \in (\mathcal{A} \otimes \mathcal{I} \mathbin{\hat{\otimes}} N)^1 \tag{2-10}$$

is a Maurer–Cartan element for $\mathcal{A} \otimes \mathcal{I}$ if and only if γ and $\tilde{\gamma}$ are Maurer–Cartan elements for $\mathcal{A}$, and h satisfies (1-2). This makes Lemma 2.4 particularly intuitive. Another possible interpretation goes as follows. Let's add a strict unit, forming $\mathbb{Z}e \oplus \mathcal{A} \mathbin{\hat{\otimes}} N$. There is an A_∞–category whose objects are Maurer–Cartan elements in $\mathcal{A} \mathbin{\hat{\otimes}} N$, with morphisms between any two elements given by $\mathbb{Z}e \oplus \mathcal{A} \mathbin{\hat{\otimes}} N$. The differential for morphisms $\tilde{\gamma} \to \gamma$ is

$$g \mapsto \sum_{p,q} \mu^{p+q+1}_{\mathbb{Z}e \oplus \mathcal{A} \hat{\otimes} N}(\overbrace{\gamma, \dots, \gamma}^{p}, g, \overbrace{\tilde{\gamma}, \dots, \tilde{\gamma}}^{q}), \tag{2-11}$$

and the formulae for higher A_∞–compositions are similar. Clearly, h satisfies (1-2) if and only if $g = e + h$ is a closed morphism $\tilde{\gamma} \to \gamma$ in our category. This viewpoint can be useful when thinking about the transitivity and functoriality of the notion of equivalence. Finally, if $\mathcal{A}$ is homologically unital, one can introduce a modified version of the Maurer–Cartan category, by setting the morphisms between objects to be $\mathcal{A} \otimes (\mathbb{Z}1 \oplus N)$, which means using the natural identity of $\mathcal{A}$ rather than artificially adjoining one. The resulting version of our previous observation (obvious in the strictly unital case, and generalized from there using Lemmas 2.2 and 2.5) is this:

Lemma 2.6 *Suppose that $\mathcal{A}$ is homologically unital. Then, two Maurer–Cartan solutions are equivalent if and only if there is a $g \in \mathcal{A}^0 \mathbin{\hat{\otimes}} (\mathbb{Z}1 \oplus N)$ which, modulo N, reduces to a cocycle homologous to $e_{\mathcal{A}}$, and which satisfies*

$$\sum_{p,q} \mu^{p+q+1}_{\mathcal{A}}(\overbrace{\gamma, \dots, \gamma}^{p}, g, \overbrace{\tilde{\gamma}, \dots, \tilde{\gamma}}^{q}) = 0. \tag{2-12}$$

2c Hochschild cochains

As before, let $\mathcal{A}$ be an A_∞–ring. Our attention will now shift to its Hochschild complex (the complex underlying Hochschild cohomology)

$$\mathcal{C} = \mathrm{CC}^*(\mathcal{A}) = \prod_{d \geq 0} \mathrm{Hom}(\mathcal{A}[1]^{\otimes d}, \mathcal{A}). \tag{2-13}$$

The Hochschild differential is

$$\begin{aligned}(d_{\mathcal{C}} c)^d(a_1, \dots, a_d) &= -\sum_{ij} (-1)^{\maltese_i \cdot \|c\|} \mu_{\mathcal{A}}^{d-j+1}(a_1, \dots, c^j(a_{i+1}, \dots, a_{i+j}), \dots, a_d) \\ &\quad + \sum_{ij} (-1)^{\maltese_i + \|c\|} c^{d-j+1}(a_1, \dots, \mu_{\mathcal{A}}^j(a_{i+1}, \dots, a_{i+j}), \dots, a_d)\end{aligned} \tag{2-14}$$

(we apologize for the double use of d as differential and as counting the number of entries); and its cohomology is the Hochschild cohomology $HH^*(\mathcal{A})$. We will also use Hochschild cohomology with coefficients in a commutative ring R, denoted by $HH^*(\mathcal{A}; R)$, which is the cohomology of $\mathrm{CC}^*(\mathcal{A}; R) = \mathcal{C} \mathbin{\hat{\otimes}} R$ — here, completion means that we take each term in (2-13) $\otimes R$ and then their product. $\mathcal{C}$ carries a canonical A_∞–structure, with $\mu_{\mathcal{C}}^1 = -d_{\mathcal{C}}$, and where the next term is

$$\begin{aligned}\mu_{\mathcal{C}}^2(c_1, c_2)^d(a_1, \dots, a_d) = \\ \sum_{\substack{i_1, j_1, i_2, j_2 \\ i_2 \geq i_1 + j_1}} (-1)^{\maltese_{i_1} \|c_1\| + \maltese_{i_2} \|c_2\|} \mu_{\mathcal{A}}^{d-j_1-j_2+2}\big(a_1, \dots, c_1^{j_1}(a_{i_1+1}, \dots, a_{i_1+j_1}), \dots, \\ c_2^{j_2}(a_{i_2+1}, \dots, a_{i_2+j_2}), \dots, a_d\big).\end{aligned} \tag{2-15}$$

The higher-order A_∞–operations follow the same pattern as $\mu_{\mathcal{C}}^2$. If $\mathcal{A}$ has a homological unit, then so does $\mathcal{C}$. One way to show that is to apply Lemma 2.3: in that situation, the restriction from the Hochschild complex of the A_∞–category to the Hochschild complex of either $\mathcal{A}$ or $\widetilde{\mathcal{A}}$ is a homotopy equivalence, allowing one to transfer properties from $\widetilde{\mathcal{A}}$ to $\mathcal{A}$ in two steps.

Note that strictly speaking, $\mathcal{C}$ does not fit into the original context for A_∞–rings, because (2-13) is not usually free. However, it is the inverse limit of chain complexes of free groups, by using the (complete decreasing) length filtration, which is compatible with the A_∞–structure. All the associated notions have to be modified to take this "pro-object" nature into account. We have already done that when defining Hochschild cohomology with coefficients, by using the completed tensor product $\mathcal{C} \mathbin{\hat{\otimes}} R$. Maurer–Cartan elements, and homotopies between such elements, will live in such completed tensor products. To prove the analogue of Lemma 2.6 for Hochschild complexes, one again uses reduction to the strictly unital case via Lemma 2.3.

The product on Hochschild cohomology induced from $\mu^2_{\mathcal{C}}$ is graded commutative. Additionally, Hochschild cohomology has a Lie bracket of degree -1. The two combine to form the structure of a Gerstenhaber algebra. When we take coefficients in a ring with $pR = 0$, let's say for concreteness $R = \mathbb{F}_p$, there is one more operation

$$\Xi_{\mathcal{A},p} \colon HH^l(\mathcal{A};\mathbb{F}_p) \to HH^{pl-(p-1)}(\mathcal{A};\mathbb{F}_p) \quad \begin{cases} \text{for odd } l \text{ if } p > 2, \\ \text{for all } l \text{ if } p = 2. \end{cases} \tag{2-16}$$

This combines with the bracket to form a restricted Lie algebra [77]. As we next explain, following [71], the underlying chain-level map can be written as a sum over trees.

Terminology 2.7 A rooted tree with d leaves is a tree which (in addition to its finite edges) has $d+1$ semi-infinite edges. One of the semi-infinite edges is singled out, and called the root; the other d are the leaves. There is a unique way of orienting edges, so that they point towards the root. Given a vertex v, write $|v|$ for its valence. Among the edges adjacent to v, there is a unique outgoing one, and $\|v\| = |v| - 1$ incoming ones.

In our applications, the rooted trees (unless otherwise indicated) come with the following structure. First, an ordering of the semi-infinite edges by $\{0,\dots,d\}$, starting with the root. Secondly, at any vertex, an ordering of the adjacent edges by $\{0,\dots,|v|\}$, again starting with the outgoing edge. A special case is that of rooted planar trees, where all orderings come from a single embedding of the tree into the plane, which implies certain compatibilities between them.

For now, we will only use rooted planar trees (the more general version will play a role later on; see Section 3b). Given such a tree and a Hochschild cochain c, one defines an operation $\mathcal{A}^{\otimes d} \to \mathcal{A}$, by starting with elements of $\mathcal{A}$ at the leaves, and having $c^{\|v\|}$ act at each vertex, with the output of that fed into the next vertex on our way to the root. To define the chain map underlying (2-16), one considers those operations for trees with p vertices, and adds them up with certain multiplicities: the multiplicity of a tree is the number of ways to order its vertices, so that the ordering increases when going towards the root ("causal orderings"). For $p = 2$, we get

$$(\Xi_{\mathcal{A},2}c)^d(a_1,\dots,a_d) = \sum_{ij} c^{d-j+1}(a_1,\dots,c^j(a_{i+1},\dots,a_{i+j}),\dots,a_d). \tag{2-17}$$

Figure 1: Sample trees that correspond to expressions from (2-18): $c^5(a_1, c^3(a_2, a_3, a_4), a_5, c^0, a_6)$, left, and $c^3(a_1, c^2(c^3(a_2, a_3, a_4), a_5), a_6)$, right.

This is usually written as $c \circ c$, where $\circ$ is the operation which underlies the homotopy commutativity of $\mu^2_{\mathcal{C}}$, and which upon antisymmetrization yields the Lie bracket. The $p = 3$ case is less familiar [71, Example 3.3]:

$$
\begin{aligned}
(\Xi_{\mathcal{A},3}c)^d(a_1, \dots, a_d) = {} & \\
2 \sum_{\substack{i_1,j_1,i_2,j_2 \\ i_1+j_1 \le i_2}} & c^{d-j_1-j_2+2}\big(a_1, \dots, c^{j_1}(a_{i_1+1}, \dots, a_{i_1+j_1}), \\
& \qquad \dots, c^{j_2}(a_{i_2+1}, \dots, a_{i_2+j_2}), \dots, a_d\big) \\
+ \sum_{i_1,j_1,i_2,j_2} & c^{d-j_1-j_2+2}\big(a_1, \dots, c^{j_1}\big(a_{i_1+1}, \dots c^{j_2}(a_{i_2+1}, \dots, a_{i_2+j_2}), \\
& \qquad \dots, a_{i_1+j_1+j_2-1}\big), \dots, a_d\big).
\end{aligned} \tag{2-18}
$$

The summands in (2-18) correspond to trees as in Figure 1, where that on the left admits two causal orderings. Koszul signs as in (2-15) are absent here, since $\|c\|$ is even; recall that for odd p, the operation $\Xi_{\mathcal{A},p}$ is only defined on odd-degree Hochschild cohomology.

Example 2.8 The first terms of $d_{\mathcal{C}}(c) = 0$, for $\|c\|$ even, are

$$
\begin{aligned}
& \mu^1_{\mathcal{A}}(c^0) = 0, \\
& \mu^1_{\mathcal{A}}(c^1(a)) + \mu^2_{\mathcal{A}}(a, c^0) + \mu^2_{\mathcal{A}}(c^0, a) = c^1(\mu^1_{\mathcal{A}}(a)), \\
& \mu^1_{\mathcal{A}}(c^2(a_1, a_2)) + \mu^2_{\mathcal{A}}(c^1(a_1), a_2) + \mu^2_{\mathcal{A}}(a_1, c^1(a_2)) \\
& \qquad + \mu^3_{\mathcal{A}}(c^0, a_1, a_2) + \mu^3_{\mathcal{A}}(a_1, c^0, a_2) + \mu^3_{\mathcal{A}}(a_1, a_2, c^0) \\
& \qquad\qquad = c^1(\mu^2_{\mathcal{A}}(a_1, a_2)) + c^2(\mu^1_{\mathcal{A}}(a_1), a_2) + (-1)^{\|a_1\|} c^2(a_1, \mu^1_{\mathcal{A}}(a_2)).
\end{aligned} \tag{2-19}
$$

The constant term in (2-18) is

$$
(\Xi_{\mathcal{A},3}c)^0 = 2c^2(c^0, c^0) + c^1(c^1(c^0)). \tag{2-20}
$$

One sees that this is again a cocycle modulo 3:

$$
\begin{aligned}
&\mu^1_{\mathcal{A}}(c^2(c^0,c^0)) \\
&\quad = -\mu^2_{\mathcal{A}}(c^1(c^0),c^0) - \mu^2_{\mathcal{A}}(c^0,c^1(c^0)) - 3\mu^3_{\mathcal{A}}(c^0,c^0,c^0) + c^1(\mu^2_{\mathcal{A}}(c^0,c^0)) \\
&\quad = \mu^1_{\mathcal{A}}(c^1(c^1(c^0))) - 3\mu^3_{\mathcal{A}}(c^0,c^0,c^0) + 3c^1(\mu^2_{\mathcal{A}}(c^0,c^0)).
\end{aligned}
\tag{2-21}
$$

Example 2.9 Suppose that $\mathcal{A}$ is a differential graded algebra, ie $\mu^d_{\mathcal{A}} = 0$ for $d \geq 3$. A derivation of $\mathcal{A}$ gives a cocycle in $\mathcal{C}^1$, and applying (2-16) amounts to taking the p^{th} iterate of that derivation.

2d The formal group structure

Given an adic ring N, let $\mathcal{C} \mathbin{\widehat{\otimes}} N$ be the space obtained by applying $\widehat{\otimes}\, N$ to each factor in (2-13) and then again forming their product. We consider Maurer–Cartan elements $\gamma \in \mathcal{C} \mathbin{\widehat{\otimes}} N$. Concretely, the first terms are

$$
\begin{aligned}
&\gamma^0 \in \mathcal{A}^1 \mathbin{\widehat{\otimes}} N, \quad \sum_d \mu^d_{\mathcal{A}}(\gamma^0,\dots,\gamma^0) = 0, \\
&\gamma^1 \in \mathrm{Hom}(\mathcal{A},\mathcal{A})^0 \mathbin{\widehat{\otimes}} N, \quad \sum_{p,q} \mu^{p+q+1}_{\mathcal{A}}(\underbrace{\gamma^0,\dots,\gamma^0}_{p}, a, \underbrace{\gamma^0 \dots, \gamma^0}_{q}) = \gamma^1(\mu^1_{\mathcal{A}}(a)),
\end{aligned}
\tag{2-22}
$$

and so on. One can think of γ as a formal deformation of the identity endomorphism of $\mathcal{A}$. What this means is that γ satisfies (1-1) if and only if, over $\mathbb{Z}1 \oplus N$,

$$
\phi^d = \begin{cases} \mathrm{id}_{\mathcal{A}} + \gamma^1 & \text{if } d = 1, \\ \gamma^d & \text{if } d \neq 1, \end{cases}
\tag{2-23}
$$

satisfies the (curved) A_∞–homomorphism equations. Similarly, two Maurer–Cartan solutions are equivalent (1-2) if the associated A_∞–homomorphisms (2-23) are (curved) homotopic. The standard composition of A_∞–homomorphisms (2-23) leads to the following composition law for Maurer–Cartan solutions:

$$
\begin{aligned}
&(\gamma_1 \bullet \gamma_2)^d(a_1,\dots,a_d) = \\
&\gamma_2^d(a_1,\dots,a_d) \\
&+ \sum_{\substack{m \geq 0 \\ i_1,j_1,\dots,i_m,j_m \\ i_1+j_1 \leq i_2,\dots, \\ i_{m-1}+j_{m-1} \leq i_m}} \gamma_1^{d-j_1-\dots-j_m+m}\bigl(a_1,\dots,\gamma_2^{j_1}(a_{i_1+1},\dots,a_{i_1+j_1}), \\
&\qquad a_{i_1+j_1+1},\dots,\gamma_2^{j_2}(a_{i_2+1},\dots,a_{i_2+j_2}), \\
&\qquad \dots,\gamma_2^{j_m}(a_{i_m+1},\dots,a_{i_m+j_m}),\dots\bigr).
\end{aligned}
\tag{2-24}
$$

This is strictly associative, and descends to a product on $\mathrm{MC}(\mathcal{C}; N)$. Moreover, by explicitly solving the equation $\phi_2\phi_1 = \mathrm{id}_{\mathcal{A}}$, one sees that this composition has inverses.

The outcome is that $N \mapsto \mathrm{MC}(\mathcal{C}; N)$ comes with the structure of a "formal group". The analogue of Theorem 1.9 in this algebraic context is [71, equation (3-1)]:

Lemma 2.10 *There is a commutative diagram*

$$\begin{tikzcd}
\mathrm{MC}(\mathcal{C}; q\mathbb{F}_p[q]/q^{p+1}) \arrow[rr, "p^{\text{th}} \text{ power of the formal group}"] \arrow[d, "\text{projection}"'] & & \mathrm{MC}(\mathcal{C}; q\mathbb{F}_p[q]/q^{p+1}) \\
\mathrm{MC}(\mathcal{C}; q\mathbb{F}_p[q]/q^2) \arrow[d, equal] & & \mathrm{MC}(\mathcal{C}; q^p\mathbb{F}_p[q]/q^{p+1}) \arrow[u, "\text{inclusion}"'] \arrow[d, equal] \\
HH^1(\mathcal{A}; \mathbb{F}_p) \arrow[rr, "\Xi_{\mathcal{A},p}"] & & HH^1(\mathcal{A}; \mathbb{F}_p)
\end{tikzcd} \tag{2-25}$$

The proof is quite straightforward. Namely, let's iterate (2-24) to form the p^{th} power of a Maurer–Cartan element γ. The outcome can be written as a sum over rooted planar trees, with multiplicities. These multiplicities count "causal labelings" of trees, where the vertices are labeled by $\{1, \dots, p\}$ and the numbers increase when going towards the root. This limits the depth of the tree to be $\leq p$, but does not by itself limit the number of vertices, since several vertices can carry the same label. However, in the formula for the p^{th} power map, each vertex carries a copy of γ, and since the coefficient ring $N = q\mathbb{F}_p[q]/q^{p+1}$ satisfies $N^{p+1} = 0$, the contribution from trees with $> p$ vertices vanishes. The labels on trees with $\leq p$ vertices can be thought of as consisting of two pieces: a choice of subset of $\{1, \dots, p\}$, and then a choice of labels which uses all numbers in that subset, and which obeys the causality condition. From that, it follows that the only trees with nontrivial mod p contribution are those with exactly p vertices, and where each label is used once. If we write $\gamma = cq + O(q^2)$, it then follows that

$$\overbrace{\gamma \bullet \cdots \bullet \gamma}^{p} = \Xi_{\mathcal{A},p}(c)\, q^p \in \mathcal{C} \mathbin{\hat{\otimes}} q\mathbb{F}_p[q]/q^{p+1}. \tag{2-26}$$

Remark 2.11 In characteristic zero, the deformation theory associated to the Maurer–Cartan equation in $\mathcal{C}$ is unobstructed: as a concrete illustration, the truncation map

$$\mathrm{MC}(\mathcal{C}; q\mathbb{Q}[\![q]\!]) \to \mathrm{MC}(\mathcal{C}; q\mathbb{Q}[q]/q^2) = HH^1(\mathcal{A}; \mathbb{Q}) \tag{2-27}$$

is onto. This is closely related to the formal group structure, since one can prove it by formal exponentiation. The analogous statement in positive characteristic is no longer generally true. The square of a class in $HH^1(\mathcal{A}; \mathbb{F}_2)$ is not necessarily zero, and that gives an obstruction to lifting to $\mathrm{MC}(\mathcal{C}; q\mathbb{F}_2[q]/q^3)$. As an example, take a polynomial ring $\mathcal{A} = \mathbb{Z}[a]$ with $|a| = 1$; the element a becomes central over $\mathbb{F}_2$, hence gives a

Hochschild cohomology class. Instead, one could look at the 2–adic lifting problem, but that's also obstructed: in the first step, which means lifting to $\mathrm{MC}(\mathcal{C}; 2\mathbb{Z}/8\mathbb{Z})$, the requirement is that the square of the Hochschild cohomology class must be equal to its Bockstein (which fails in the same example).

Remark 2.12 If $\mathcal{A}$ is a dg algebra, the Hochschild complex has the same structure. Let's follow classical notation and write $\smile$ for the product on Hochschild cochains. The Maurer–Cartan equation is

$$d_{\mathcal{C}}\gamma + \gamma \smile \gamma = 0, \tag{2-28}$$

and two solutions are equivalent if

$$d_{\mathcal{C}}h + (\gamma + \gamma \smile h) - (\tilde{\gamma} + h \smile \tilde{\gamma}) = 0. \tag{2-29}$$

The composition law (2-24) can be written in terms of the brace operations from [72] as

$$\gamma_1 \bullet \gamma_2 = \gamma_2 + \sum_{m \geq 0} \gamma_1\{\overbrace{\gamma_2, \ldots, \gamma_2}^{m}\}. \tag{2-30}$$

When put in this way, the formalism can be generalized to any complex $\mathcal{C}$ which is an algebra over the braces operad [53], since that exactly provides the operations used in (2-28)–(2-30). The formula (2-30) can be viewed as an application of a construction [27] (see [76] for a review and further context) which equips the tensor coalgebra of $\mathcal{C}$ with a bialgebra structure. It is possible that the geometric results in this paper could be similarly sharpened, replacing "formal groups" with a suitable bialgebra language (where the comultiplication would be the standard tensor coalgebra structure, but the multiplication would be A_∞); however, that would likely require the full generality of Bottman's witch ball spaces.

3 Parameter spaces

This section discusses the moduli spaces underlying our constructions. This is mostly an exposition of known material; the small amount that may be new appears towards the end of the section. Stasheff associahedra, Deligne–Mumford spaces, and Fulton–MacPherson spaces (for the latter, originally in their homotopy equivalent guise [62] as the little squares operad) belong to classical algebraic topology and geometry, and we include a brief exposition mainly as a warmup exercise. The more complicated spaces are borrowed from the theory of Lagrangian correspondences, variously combining [49; 48; 9; 22; 5; 6].

3a Associahedra

The Stasheff spaces (associahedra) S_d, for $d \geq 2$, are compactifications of the space of ordered point configurations on the real line, modulo translations and positive dilations, meaning of

$$\frac{\{(s_1,\dots,s_d)\in\mathbb{R}^d : s_1<\cdots<s_d\}}{\{(s_1,\dots,s_d)\sim(\tau(s_1),\dots,\tau(s_d)) \text{ for } \tau(s)=\lambda+\mu s \text{ with } \lambda\in\mathbb{R},\ \mu>0\}}. \tag{3-1}$$

The collection $\{S_d\}$ has the structure of a nonsymmetric operad, given by maps

$$\prod_v S_{\|v\|} \xrightarrow{T} S_d, \tag{3-2}$$

one for each rooted planar tree T with $d+1$ semi-infinite edges, and where every vertex v has valence $|v| \geq 3$ (see Terminology 2.7; it will be our standard procedure to just denote such maps by the underlying tree). The single-vertex tree is a trivial special case, since it gives rise to the identity map on S_d.

Topologically, S_d is a (contractible) compact manifold with boundary, whose interior is (3-1), and whose boundary is the union of the images of the nontrivial maps (3-2). One can get a slightly more precise description by introducing a suitable smooth structure, for instance by embedding the Stasheff spaces into the real locus of Deligne–Mumford spaces. Then S_d becomes a smooth (and in fact real subanalytic) manifold with corners, whose open strata are the images of $\prod_v (S_{\|v\|} \setminus \partial S_{\|v\|})$ under (3-2).

We orient S_d by picking, on the interior (3-1), the parametrization where (s_1, s_2) are fixed, and using the standard orientation of the remaining parameters $(s_3, \dots, s_d)$.

3b Fulton–MacPherson spaces

The Fulton–MacPherson spaces[2] FM_d, for $d \geq 2$, are compactifications of planar configuration space up to translations and positive dilations:

$$\frac{\{(z_1,\dots,z_d)\in\mathbb{C}^d : z_i\neq z_j \text{ for } i\neq j\}}{\{(z_1,\dots,z_d)\sim(\tau(z_1),\dots,\tau(z_d)) \text{ for } \tau(z)=\lambda+\mu z \text{ with } \lambda\in\mathbb{C},\ \mu>0\}}. \tag{3-3}$$

The (symmetric) operad structure on $\{\mathrm{FM}_d\}$ comes from permutations of the z_k, together with maps similar to (3-2),

$$\prod_v \mathrm{FM}_{\|v\|} \xrightarrow{T} \mathrm{FM}_d. \tag{3-4}$$

[2]The terminology is taken from [29]; versions of the construction arose in [3; 25; 37]

Here, the rooted trees T come with our usual structure (see Terminology 2.7), but are not necessarily planar. Changing the ordering of the semi-infinite edges of T amounts to composing (3-4) with an element of Sym_d on the left; and changing the orderings at the vertices amounts to composing (3-4) with an element of $\prod_v \mathrm{Sym}_{\|v\|}$ on the right. The inclusion $\mathbb{R} \subset \mathbb{C}$ induces maps

$$S_d \to \mathrm{FM}_d, \tag{3-5}$$

which are compatible with (3-2) and (3-4) (they form a morphism of nonsymmetric operads).

As before, FM_d is topologically a compact manifold with boundary. One can complexify it by considering point configurations in $\mathbb{C}^2$, which yields a smooth compact complex manifold, and then embed FM_d into the real locus of that. As a consequence, it inherits the structure of a smooth (or real subanalytic) manifold with corners, just as in the case of the associahedra.

To orient FM_d, we consider representatives in (3-3) where z_1 and $|z_1 - z_2|$ are fixed. Then, rotating z_2 anticlockwise around z_1 yields the first coordinate, and the remaining coordinates are $(z_3, \dots, z_d)$ with their complex orientations. Equivalently, consider the classical configuration space $\mathrm{Conf}_d(\mathbb{C})$, of which (3-3) is a quotient by the action of $(\lambda, \mu) \in \mathbb{C} \times \mathbb{R}^{>0}$. The Lie algebra of that group fits into an exact sequence

$$0 \to \mathbb{C} \oplus \mathbb{R} \to T_{(z_1,\dots,z_d)} \mathrm{Conf}_d(\mathbb{C}) \to T_{(z_1,\dots,z_d)} \mathrm{FM}_d \to 0; \tag{3-6}$$

our orientation of the quotient is compatible with that sequence and with the complex orientation of $\mathrm{Conf}_d(\mathbb{C})$. In particular, Sym_d acts orientation-preservingly.

3c Deligne–Mumford spaces

For most of this paper, we will write DM_d for the Deligne–Mumford moduli space of genus 0 curves with $d+1$ marked points, bringing it in line with the notation for the other moduli spaces. One can consider it as a compactification of

$$\frac{\{(z_1, \dots, z_d) \in \mathbb{C}^d : z_i \neq z_j \text{ for } i \neq j\}}{\{(z_1, \dots, z_d) \sim (\tau(z_1), \dots, \tau(z_d)) \text{ for } \tau(z) = \lambda + \mu z \text{ with } \lambda \in \mathbb{C},\, \mu \in \mathbb{C}^*\}}, \tag{3-7}$$

which is a free S^1–quotient of (3-3). The operadic structure takes on exactly the same form as for Fulton–MacPherson spaces. Indeed, the quotient map on configuration spaces extends to a map

$$\mathrm{FM}_d \to \mathrm{DM}_d, \tag{3-8}$$

which is compatible with (3-4) and its Deligne–Mumford counterpart. We adopt the usual orientation of DM_d as a complex manifold.

3d Colored multiplihedra

Ma'u, Wehrheim and Woodward [49; 48] introduced a geometric interpretation of the classical multiplihedra, as well as certain generalizations. We will call these spaces colored multiplihedra, and denote them by

$$\mathrm{MWW}_{d_1,\dots,d_r} \quad \text{for } r \geq 1,\ d_1,\dots,d_r \geq 0,\ d = d_1+\cdots+d_r > 0. \tag{3-9}$$

They are compactifications of

$$\frac{\{(s_{1,1},\dots,s_{1,d_1};\dots;s_{r,1},\dots,s_{r,d_r}) \in \mathbb{R}^d : s_{k,1} < \cdots < s_{k,d_k} \text{ for each } k\}}{\{s_{k,i} \sim s_{k,i}+\mu \text{ for } \mu \in \mathbb{R}\}}. \tag{3-10}$$

The intuitive meaning of (3-10) is that we have d points on the real line, which are divided into r colors, with d_k points of any given color k. Points of different colors can have the same position, while those of the same color are distinct and lie on the real line in increasing order. We denote the compactification by $\mathrm{MWW}_{d_1,\dots,d_r}$. It tracks what happens on a large scale, meaning the relative speeds as points diverge from each other, as well as on the small scale, where points of the same color converge. Therefore, a point in the compactification consists of "screens" (terminology taken from [25]) which are either "large-scale", "mid-scale", or "small-scale". Correspondingly, the analogue of (3-2) is of the form

$$\prod_{v \text{ large}} S_{\|v\|} \times \prod_{v \text{ mid}} \mathrm{MWW}_{\|v\|_1,\dots,\|v\|_r} \times \prod_{v \text{ small}} S_{\|v\|} \xrightarrow{T} \mathrm{MWW}_{d_1,\dots,d_r}. \tag{3-11}$$

Here, the tree T has $d+1$ semi-infinite edges. We still single out a root, but the leaves are now divided into subsets of orders d_k, each subset being then ordered by $\{1,\dots,d_k\}$. Each vertex has one of three scales. The mid-scale vertices have the same kind of combinatorial data attached to them as the entire tree: their incoming edges are divided into r subsets of different colors, whose sizes we denote by $\|v\|_1,\dots,\|v\|_r$, and then ordered within each subset. The large-scale vertices and small-scale vertices just come with an ordering of the incoming edges. The small-scale vertices are also labeled with a color in $\{1,\dots,r\}$. Any path going from a leaf of color k to the root travels in nondecreasing order of scale: first through any number (which can be zero) of small-scale vertices of color k; then through exactly one single mid-scale vertex, which it enters by an edge with color k; and finally, through any number (which can be zero)

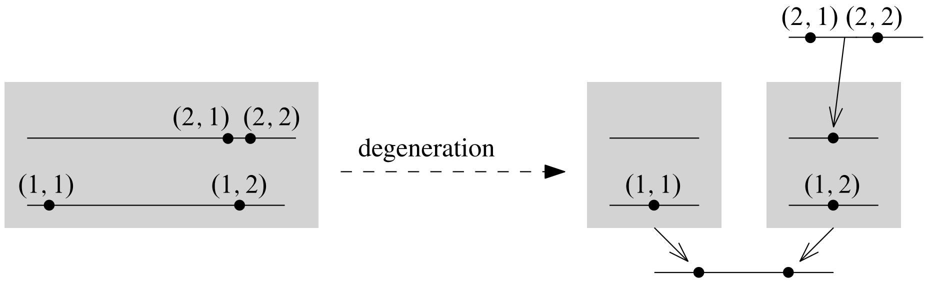

Figure 2: A degeneration in $\mathrm{MWW}_{2,2}$; see Example 3.1. The shaded regions are "mid-scale screens"; we have drawn marked points of different colors as lying on separate copies of the real line.

of large-scale vertices. There are compatibility conditions between the orderings, which are somewhat tedious to write down combinatorially; see [48, Section 6] — they are similar in principle to those for planar rooted trees, but concern each color separately.

Example 3.1 Suppose that in $\mathrm{MWW}_{2,2}$ we have a sequence of configurations where one point (of the first color) moves to $-\infty$, and the remaining three points move towards the same position. The outcome is shown in Figure 2.

Topologically, $\mathrm{MWW}_{d_1,\dots,d_r}$ is again a compact manifold with boundary, having (3-10) as its interior. Note that the codimension of the image of (3-4) is the number of small-scale plus large-scale vertices, mid-scale vertices being irrelevant. As a consequence of the resulting combinatorial structure of boundary strata, $\mathrm{MWW}_{d_1,\dots,d_r}$ can't be made into a smooth manifold with corners in the same way as the previously considered moduli spaces. However, it is naturally a (subanalytic) manifold with generalized corners in the sense of [36]. To prove that, one introduces a complexification as in [49; 8], which is a complex variety with toric singularities, and embeds $\mathrm{MWW}_{d_1,\dots,d_r}$ into its real locus.

As for orientations, we orient (3-10) by ordering the coordinates lexicographically, and then keeping the first one fixed to break the translation-invariance.

Example 3.2 In the spaces $\mathrm{MWW}_{1,\dots,1}$, no small-scale vertices can appear. The maps (3-11) with zero-dimensional domains correspond to trivalent planar rooted trees with an additional ordering of the r leaves, hence there are $(2r-2)!/(r-1)!$ of them. For instance, the two-dimensional space $\mathrm{MWW}_{1,1,1}$ is a 12–gon; see Figure 3. The boundary sides each have either one large-scale screen containing three points, or one

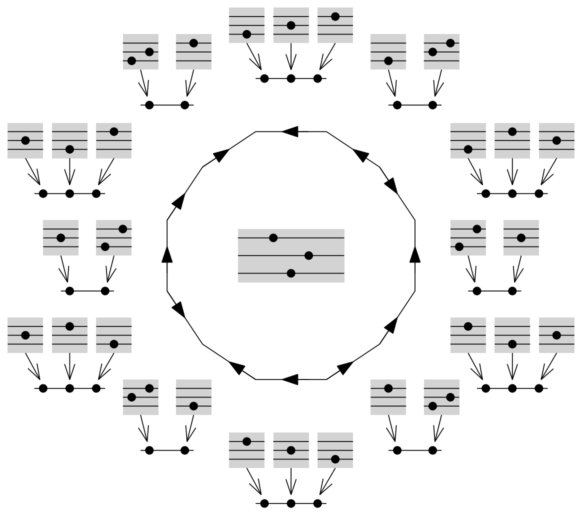

Figure 3: The space $\mathrm{MWW}_{1,1,1}$ from Example 3.2.

mid-scale screen with two points (each possibility occurs six times). Figure 4 shows in more detail a neighborhood of one of the corners of the 12–gon, and in particular, the degenerate configuration associated to the vertex.

Example 3.3 The space $\mathrm{MWW}_{2,1}$ is an octagon; see Figure 5. There are only two points which belong to the same color, hence only one way in which a small-scale vertex can occur, which is the boundary side at the top of our figure. The other boundary sides are of two kinds, as in Example 3.2.

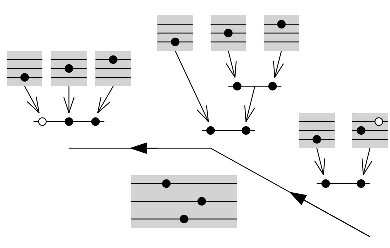

Figure 4: A specific part of $\mathrm{MWW}_{1,1,1}$, compare Figure 3. As one approaches the vertex along the edge from the left, the leftmost of the three points on the large-scale screen moves to $-\infty$. As one approaches it along the other edge from the right, the rightmost point on the mid-scale screens moves to $+\infty$. We have colored the points that we think of as moving white. (Of course, because of translation invariance, there are other equivalent ways of thinking about the degenerations.)

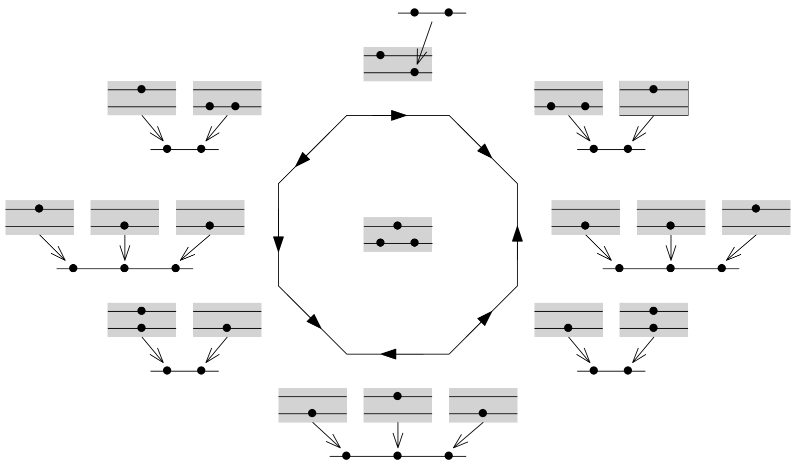

Figure 5: The space $\mathrm{MWW}_{2,1}$ from Example 3.3.

One can associate to a real colored configuration a complex configuration, by setting

$$z_{k,i} = s_{k,i} + k\sqrt{-1} \tag{3-12}$$

and then ordering the $z_{k,i}$ lexicographically (we use $\sqrt{-1}$ here to avoid notational confusion with the index i). This extends to a continuous map

$$\mathrm{MWW}_{d_1,\dots,d_r} \to \mathrm{FM}_d, \quad \text{provided that } d = d_1 + \dots + d_r \geq 2. \tag{3-13}$$

In terms of (3-11), the extension uses the same formula (3-12) for the points on each mid-scale screen, while the small-scale and large-scale screens use (3-5). To be precise, there is one exception: mid-scale vertices with $|v| = 2$ have no Fulton–MacPherson counterpart, and we simply forget about them, which is unproblematic since $\mathrm{MWW}_{0,\dots,0,1,0,\dots,0} = \text{point}$. There is a commutative diagram involving (3-13) as well as (3-5), (3-4), (3-11):

(3-14)
$$\begin{tikzcd}
\prod_{v \text{ small}} S_{\|v\|} \times \prod_{v \text{ mid}} \mathrm{MWW}_{\|v\|_1,\dots,\|v\|_r} \times \prod_{v \text{ large}} S_{\|v\|} \arrow[r, "(3\text{-}11)"] \arrow[d, "(3\text{-}13)"'] & \mathrm{MWW}_{d_1,\dots,d_r} \arrow[dd, "(3\text{-}13)"] \\
\prod_{v \text{ small}} S_{\|v\|} \times \prod_{v \text{ mid}, |v|>2} \mathrm{FM}_{\|v\|} \times \prod_{v \text{ large}} S_{\|v\|} \arrow[d, "(3\text{-}5)"'] & \\
\prod_{|v|>2} \mathrm{FM}_{\|v\|} \arrow[r, "(3\text{-}4)"'] & \mathrm{FM}_d
\end{tikzcd}$$

It can be convenient to allow more flexibility in the construction of (3-13). Namely, suppose that we have a collection of continuous functions

$$\tau_{d_1,\dots,d_r} = (\tau_{d_1,\dots,d_r,1,1},\dots,\tau_{d_1,\dots,d_r,r,d_r})\colon \mathrm{MWW}_{d_1,\dots,d_r} \to \mathbb{R}^d \tag{3-15}$$

with the following properties. In the interior of our space,

(3-16) $\tau_{d_1,\dots,d_r,k,i} < \tau_{d_1,\dots,d_r,l,j}$ at any point of (3-10) where $s_{k,i} = s_{l,j}$ for some $k < l$ and i, j.

Take the pullback of $\tau_{d_1,\dots,d_r}$ by (3-11) for some tree T. Each index (k,i) corresponds to a leaf of T, and the path from that leaf to the root enters a single mid-scale vertex v through an incoming edge labeled (k, j). Then, we require that the (k,i) component of the pullback be given by the (k, j)–component of $\tau_{\|v\|_1,\dots,\|v\|_r}$, as a function on the product in (3-11). Instead of (3-12), we can then set

$$z_{k,i} = s_{k,i} + \tau_{d_1,\dots,d_r,k,i}\sqrt{-1}. \tag{3-17}$$

Intuitively, the imaginary parts of the $z_{k,i}$ can vary depending on the modular parameters, but if two points of different colors $k < l$ come to lie on the same vertical axis, the point with the higher color l always passes above that of color k (in contrast, points of the same color still collide, "bubbling off" into a small-scale screen). The consistency condition we have imposed on (3-15) ensures that (3-17) extends to a continuous map (3-13), with the same boundary compatibilities (3-14) as before. This is a strict generalization of the previous construction, since the constant functions $\tau_{d_1,\dots,d_r,k,i} = k$ clearly satisfy our conditions. More general choices of (3-15) can be defined inductively by extension from the boundary of $\mathrm{MWW}_{d_1,\dots,d_r}$ to the entire space, which is unproblematic since (3-16) is a convex condition.

As one application of (3-17), note that we have (orientation-preserving) identifications

$$\mathrm{MWW}_{d_1,\dots,d_r} = \mathrm{MWW}_{d_1,\dots,d_{l-1},d_{l+1},\dots,d_r} \quad \text{if } d_l = 0. \tag{3-18}$$

According to the original formula (3-12), these two isomorphic spaces come with different maps to FM_d. However, when constructing the functions (3-15), one can additionally achieve that

$$\tau_{d_1,\dots,d_r,k,i} = \begin{cases} \tau_{d_1,\dots,d_{l-1},d_{l+1},\dots,k,i} & \text{for } k < l, \\ \tau_{d_1,\dots,d_{l-1},d_{l+1},\dots,k-1,i} & \text{for } k > l \end{cases} \qquad \text{if } d_l = 0, \tag{3-19}$$

and then the maps (3-13) obtained from (3-17) become compatible with (3-18).

3e Witch ball spaces

Our next topic is a simplified version of Bottman's witch ball spaces [6], for didactic reasons: we won't use them as such, but the discussion serves as a preparation for a related construction to be carried out afterwards. Our notation is

$$B_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}, \quad \text{where } r \geq 2,\ d = d_1 + \cdots + d_r > 0,\ 1 \leq m \leq r-1. \tag{3-20}$$

The interior is the configuration space (3-10) with an additional parameter $t \in (0,1)$. This parameter extends to a map

$$B_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r} \to [0,1]. \tag{3-21}$$

Over $t \in (0,1]$, we just have a copy of $(0,1] \times \mathrm{MWW}_{d_1,\dots,d_r}$. In particular, by looking at $t = 1$ one gets boundary strata inherited from (3-11), which are images of maps

$$\prod_{v \text{ large}} S_{\|v\|} \times \prod_{v \text{ mid}} \mathrm{MWW}_{\|v\|_1,\dots,\|v\|_r} \times \prod_{v \text{ small}} S_{\|v\|} \xrightarrow{T} B_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}. \tag{3-22}$$

At $t = 0$, the m^{th} and $(m+1)^{\text{st}}$ color "collide". There, the analogue of (3-22) is

$$\prod_{v \text{ mid}} \mathrm{MWW}_{\|v\|_1,\dots,\|v\|_{r-1}} \times \prod_{v \text{ small-mid}} \mathrm{MWW}_{\|v\|_1,\|v\|_2} \times \prod_{v \text{ of any other scale}} S_{\|v\|} \xrightarrow{T} B_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}. \tag{3-23}$$

This time six different scales are involved, which we (unimaginatively) call "large", "mid", "small", "small-large", "small-mid" and "small-small". Suppose that we have a path from a leaf to the root. As usual, the leaves carry colors $\{1,\dots,r\}$. If the color of our leaf is $\neq m, m+1$, things proceed as for the MWW spaces, with the path going through any number of small vertices, one mid-scale vertex, and then any number of large vertices. (There is a relabeling rule: if the color is $k > m+1$, it enters the mid-scale vertex through an edge with color $k-1$.) If the color is m (resp. $m+1$), the path first goes through small-small vertices, and then through exactly one small-mid vertex, which it enters through an edge colored by 1 (resp. 2). It then proceeds through an arbitrary number of small-large vertices, then through a mid-scale vertex, which it always enters through the m^{th} color, following by large-scale vertices. To compute the codimension of (3-23) one counts the number of "other scale" screens. Finally, our space has boundary strata which lie over the entire interval $[0,1]$, and those are images of maps

$$\prod_{v \text{ large}} S_{\|v\|} \times \prod_{v \text{ mid}}^{[0,1]} B_{\|v\|_1,\dots,(\|v\|_m,\|v\|_{m+1}),\dots,\|v\|_r} \times \prod_{v \text{ small}} S_{\|v\|} \xrightarrow{T} B_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}, \tag{3-24}$$

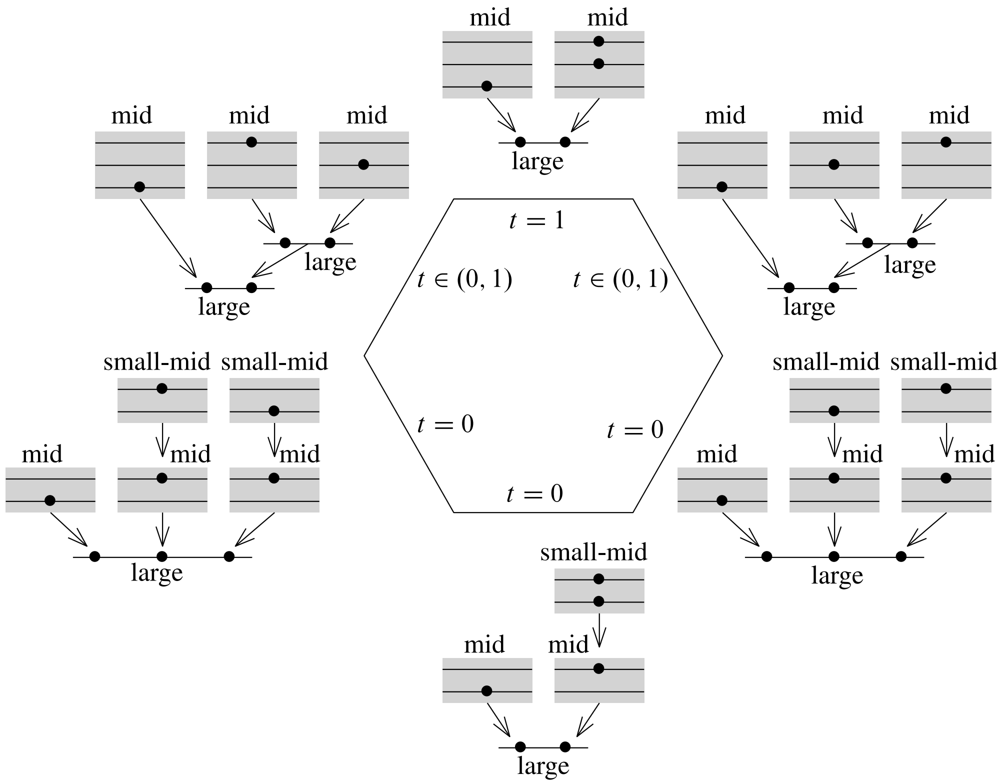

Figure 6: One of the boundary faces of $B_{1,(1,1)}$; see Example 3.4.

where the superscript means that instead of a product, we have a fiber product over (3-21); compare [7, equation (1)]. We refer to [6; 5; 7] for a detailed discussion; the results obtained there can easily be carried over to our version.

Example 3.4 The space $B_{1,(1,1)}$ is half (sliced through horizontally) of [7, Figure 1b]. Figure 6 shows one of its boundary faces of type (3-24), namely

$$S_2 \times B_{1,(0,0)} \times_{[0,1]} B_{0,(1,1)} \cong B_{(1,1)}.$$

The spaces (3-20) are topological manifolds with boundary, and smooth manifolds with generalized corners. For Bottman's witch ball spaces, this is proved in [8], and the same arguments apply to the (comparatively simpler) situation here.

As was the case for the MWW spaces, one can map our spaces to Fulton–MacPherson spaces

(3-25) $$B_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r} \to \mathrm{FM}_d \quad \text{provided that } d \geq 2,$$

compatibly with (3-22), (3-23) and (3-24). Suppose for simplicity that the maps (3-13) have been defined using (3-12). The corresponding formula for $B_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}$ is then

$$z_{k,i} = \begin{cases} s_{k,i} + k\sqrt{-1} & \text{if } k \le m, \\ s_{k,i} + (k-1+t)\sqrt{-1} & \text{if } k > m. \end{cases} \tag{3-26}$$

As before, the extension of this map to the entire space forgets any screens (necessarily of mid-scale or small-mid-scale) which carry configurations consisting only of one point.

3f Strip-shrinking spaces

We will now introduce a modification of the idea of witch ball spaces, designed to avoid the kind of fiber products which appeared in (3-24). This is inspired by [9], and correspondingly called strip-shrinking spaces. We will denote them by

$$\mathrm{SS}_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}, \quad \text{with } r \ge 2,\ d = d_1 + \dots + d_r \ge 0,\ 1 \le m \le r-1. \tag{3-27}$$

(Note that this time, unlike the situation in (3-20), it is possible to have all $d_k = 0$.) The SS spaces compactify colored configuration space as in (3-10), but without dividing by common translation. The important point is an asymmetry between the two ways in which points in the configuration can go to infinity. In the $s \to -\infty$ direction, we dictate a fairly standard behavior, where MWW spaces with r colors appear. In the $s \to +\infty$ limit, we think of the m^{th} and $(m+1)^{\text{st}}$ colored points as lying on lines that become asymptotically close to each other, at a rate of $1/s$. One way to make this more concrete is to consider the analogue of (3-26), which associates to a real configuration a complex one. Choose a function $\psi : \mathbb{R} \to (-1, 0]$ with asymptotics

$$\psi(s) \approx \begin{cases} 0 & \text{for } s \ll 0, \\ -1 + 1/s & \text{for } s \gg 0. \end{cases} \tag{3-28}$$

Then set (see Figure 7)

$$z_{k,i} = \begin{cases} s_{k,i} + k\sqrt{-1} & \text{for } k \le m, \\ s_{k,i} = s_{k,i} + (k + \psi(s_{k,i}))\sqrt{-1} & \text{for } k > m. \end{cases} \tag{3-29}$$

To relate the spaces to Fulton–MacPherson spaces, we can add two auxiliary marked points, say

$$z_{\pm} = \pm 1 + (r+1)\sqrt{-1}, \tag{3-30}$$

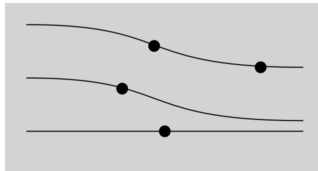

Figure 7: A point in $\mathrm{SS}_{(1,1),2}$, thought of as a configuration in $\mathbb{C}$ as in (3-29).

which stabilize the situation and otherwise stay out of the way. This gives continuous maps

$$\mathrm{SS}_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r} \to \mathrm{FM}_{d+2}. \tag{3-31}$$

For a more precise picture, consider the analogue of (3-2),

$$\begin{aligned}\prod_{v<v_* \text{ mid}} \mathrm{MWW}_{\|v\|_1,\dots,\|v\|_r} \times \mathrm{SS}_{\|v_*\|_1,\dots,\|v_*\|_r} \times \prod_{v>v_* \text{ mid}} \mathrm{MWW}_{\|v\|_1,\dots,\|v\|_{r-1}} \\ \times \prod_{v \text{ small-mid}} \mathrm{MWW}_{\|v\|_1,\|v\|_2} \times \prod_{v \text{ of any other scale}} S_{\|v\|} \\ \xrightarrow{T} \mathrm{SS}_{d_1,\dots,d_r}.\end{aligned} \tag{3-32}$$

Here, we have the same six scales as in (3-23), but with different roles. There is a distinguished mid-scale vertex, denoted by v_*, to which corresponds an SS space. All other mid-scale vertices carry MWW spaces, in two different versions: if $v < v_*$ (with respect to the ordering of mid-scale vertices determined by the ordering of the incoming edges at large-scale vertices) that space has r colors, but for $v > v_*$ there are only $r-1$ colors. The part of the tree lying on top of $v \leq v_*$ vertices consists of small-scale vertices as in (3-11), and the same is true for $v > v_*$ if the color is $\neq m$. For that remaining color, we have a structure of small-large, small-mid and small-small vertices, parallel to (3-23).

Example 3.5 Take the two-dimensional space $\mathrm{SS}_{(1,1)}$, denoting points in its interior by $(s_1; s_2)$ for brevity. Consider sequences

$$(s_1^{(k)}; s_2^{(k)}),\ k = 1, 2, \dots, \quad \text{where } s_1^{(k)} < s_2^{(k)} \text{ and } s_1^{(k)}, s_2^{(k)} \to +\infty. \tag{3-33}$$

The possible limit configurations, shown in Figure 8, correspond to the following behaviors:

(i) $(s_2^{(k)} - s_1^{(k)})/s_1^{(k)} \to +\infty$.

(ii) $(s_2^{(k)} - s_1^{(k)})/s_1^{(k)}$ converges to a nonzero constant.

(iii) $s_2^{(k)} - s_1^{(k)} \to +\infty$, but $(s_2^{(k)} - s_1^{(k)})/s_1^{(k)} \to 0$.

(iv) $s_2^{(k)} - s_1^{(k)}$ converges to a nonzero constant.

(v) $s_2^{(k)} - s_1^{(k)} \to 0$, but $s_1^{(k)}(s_2^{(k)} - s_1^{(k)}) \to +\infty$. Since the two points get increasingly close to each other, the additional mid-scale screen carries only one marked point. On the other hand, rescaling by $s_1^{(k)}$ separates the two points in the limit. This leads to the appearance of a small-large screen.

(vi) $s_2^{(k)} - s_1^{(k)} \to 0$, but $s_1^{(k)}(s_2^{(k)} - s_1^{(k)})$ converges to a constant (which can be zero). In that case, we get a small-mid scale screen with two marked points on it.

The whole space is a 14–gon (Figure 9), with three adjacent sides corresponding to (ii), (iv) and (vi) above, and corners corresponding to (i), (iii) and (iv).

The structure of SS as a compact topological space is relatively straightforward to obtain, following the model of [6]. It turns out that it is also a topological manifold with boundary, and in fact a differentiable manifold with generalized corners. The last-mentioned property deserves some discussion, since the required construction of coordinate charts, which borrows ideas from [8], is instructive in its own right.

A boundary point in $\mathit{SS}_{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}$ is given by a tree T and associated screens carrying point configurations, as in (3-32). The gluing process which associates to this point a chart in the interior involves (small) gluing parameters $\lambda_e > 0$ for the finite edges of T, subject to constraints. Our main interest lies in those constraints, but let's first recall how to think of such gluing processes. This is made slightly more complicated in our case by the fact that the screens have different natures: the vertex v_* carries a configuration of real numbers, without dividing by any group action; the mid-scale and small-mid scale vertices carry configurations which are given up to translation; and at all other scales we have configurations up to translation and rescaling. To deal with that, it is convenient to stabilize the configuration associated to the distinguished mid-scale vertex v_* by adding two points $s_\pm = \pm 1$, thought of as belonging to their own new color, just as in (3-30). To glue the screens together, we first choose specific representatives for those configurations, which are defined only up to ambiguities. Then, given any finite edge e of the tree, we take the screen associated to its source vertex, rescale the points in that configuration by λ_e, and then insert that into the target vertex by adding the real number that corresponds to the point where our configuration is being glued in. (In abbreviated notation, gluing s with scale λ into a screen at point r results in $r + \lambda s$.) After we have done that for all edges, we translate and rescale the resulting configuration to bring the points $s_\pm$ back to their original position (and then forget about those points).

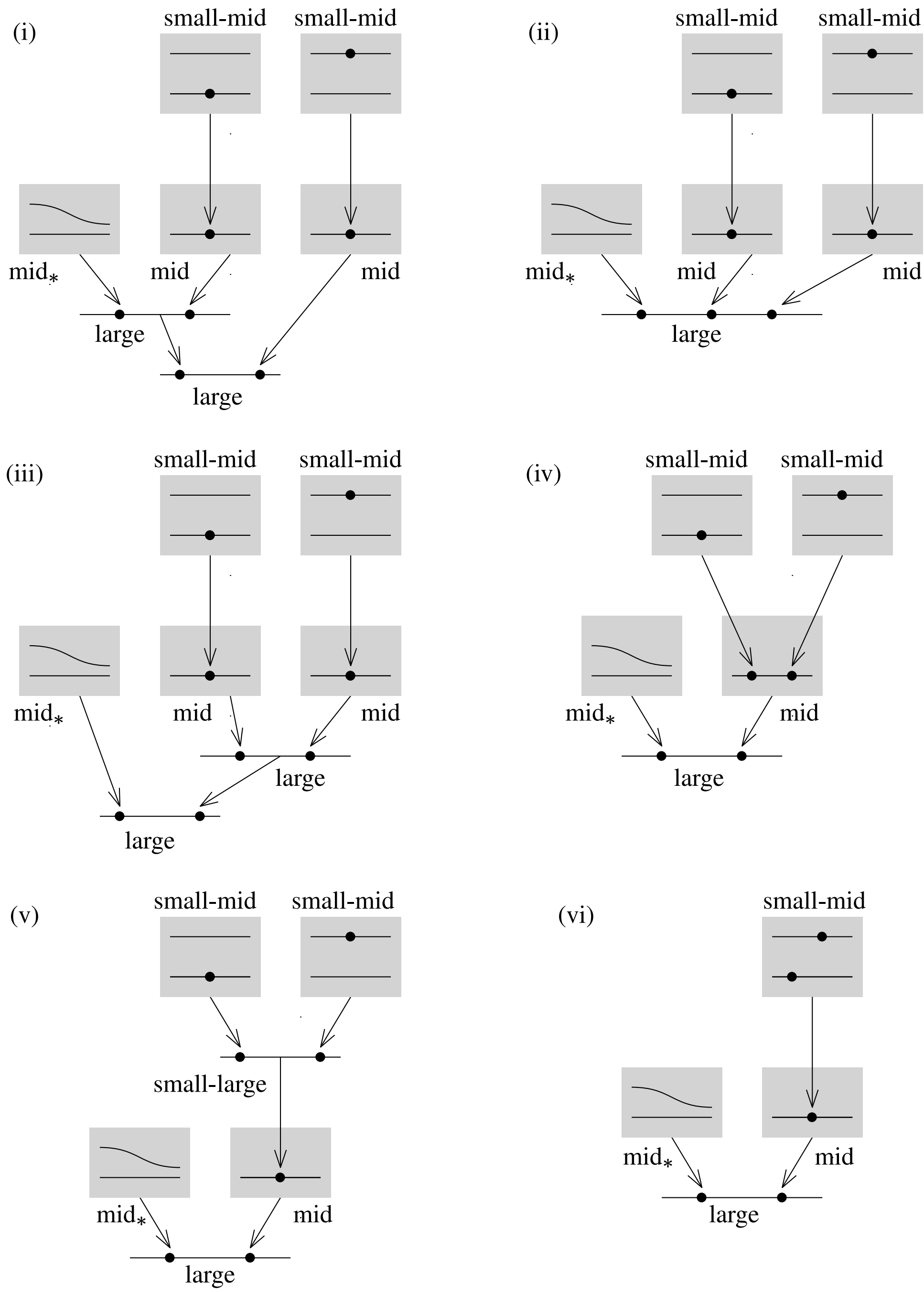

Figure 8: Some limits in $SS_{(1,1)}$, as discussed in Example 3.5. The $*$ marks the distinguished mid-scale screen.

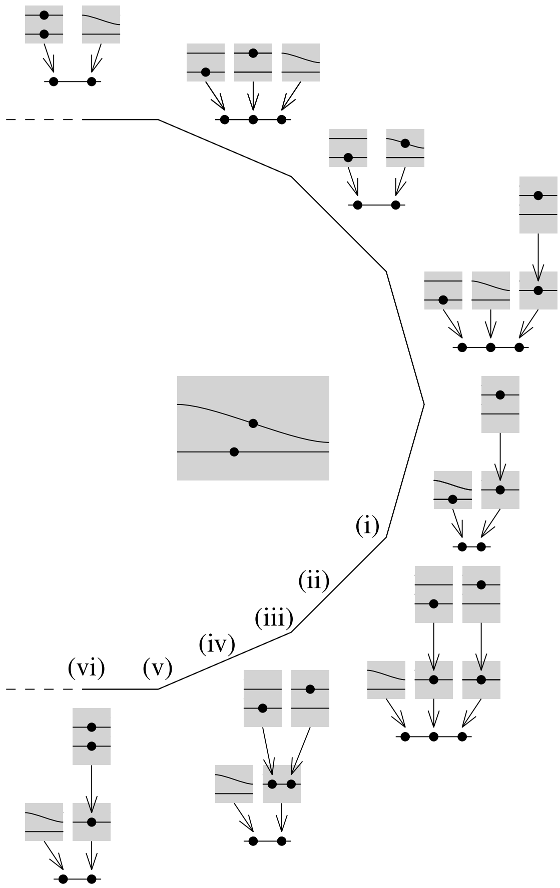

Figure 9: The space $\mathrm{SS}_{(1,1)}$ is a 14–gon; see Example 3.5. For space reasons, we have only drawn half of it. However, the drawing is arranged so that exchanging the first two colors corresponds to reflection along the vertical axis, and the missing half can be inferred from that.

It may strike the reader that there are too many gluing parameters with respect to the codimension of the boundary strata; and indeed, the parameters are not independent, but subject to constraints. To formulate those, we can think in terms of the scales that the screens acquire after gluing. For any vertex v, let λ_v be the product of the $\lambda_e^{\pm 1}$ along a path going from v_* to v, with the sign $+1$ if the path follows the edge orientation, and -1 otherwise. We also need the following terminology:

(3-34) Given a small-mid scale vertex v_-, we say that a large scale vertex v_+ is a turning point for v_- if there is a path from v_* to v_- which follows the orientation until it hits v_+, and then goes against the orientation to v_-.

For any v_- there is a unique turning point v_+. With that at hand, the relations are:

(3-35) If v is a mid-scale vertex, $\lambda_v = 1$ (this is automatic for $v = v_*$).

(3-36) If v_+ is a turning point for v_- (so v_- is small-mid scale), $\lambda_{v_+}\lambda_{v_-} = 1$.

It is easy to see that for a codimension-one stratum, all the λ_e therefore end up being the same.

Example 3.6 Consider gluing from the horizontal boundary edge at the top of Figure 9. Let's say that the large screen carries the configuration (r_1, r_2); the mid-scale screen on the left carries a configuration $(s_1; s_2)$; the remaining screen, corresponding to v_*, is empty, but we add a third color and its points $s_\pm$ as explained above. The constraint (3-35) says that the gluing parameters for both edges must be equal, so we effectively have a single parameter λ. In a first step, gluing with that parameter yields the configuration

(3-37) $$(r_1 + \lambda s_1; r_1 + \lambda s_2; r_2 - \lambda, r_2 + \lambda).$$

After that, we apply translation and rescaling which maps $r_2 \pm \lambda$ back to ± 1, that being $s \mapsto \lambda^{-1}(s - r_2)$; and (forgetting those points) we end up with

(3-38) $$(\lambda^{-1}(r_1 - r_2) + s_1; \lambda^{-1}(r_1 - r_2) + s_2) \in SS_{(1,1)} \setminus \partial SS_{(1,1)}.$$

This means that the gluing takes place in a way which preserves the size of the mid-scale screens, even though that has been obscured a bit by writing it as rescaling by λ and then its inverse.

Example 3.7 Consider the situation of the horizontal boundary edge at the bottom of Figure 9, which is also Figure 8(vi). Let's say that λ_1 and λ_2 are the gluing parameters for the edges leading to the large-scale vertices, and λ_3 that for the remaining edge. Then, (3-36) says that $\lambda_1\lambda_2^{-1} = 1$, and (3-36) that $\lambda_{v_+}\lambda_{v_-} = \lambda_1(\lambda_1\lambda_2^{-1}\lambda_3^{-1}) = \lambda_1^2\lambda_2^{-1}\lambda_3^{-1} = 1$. As mentioned before, the end result is again that all gluing parameters are equal. Suppose that the large-scale screen carries (r_1, r_2), the mid-scale screen carries r, and the small-mid scale screen carries $(s_1; s_2)$. The analogue of (3-37) is

(3-39) $$(r_2 + \lambda^2 s_1; r_2 + \lambda^2 s_2; r_1 - \lambda, r_1 + \lambda),$$

and that of (3-38) is obtained by applying $s \mapsto \lambda^{-1}(s - r_1)$, which gives

(3-40) $$(\lambda^{-1}(r_2 - r_1) + \lambda s_1; \lambda^{-1}(r_2 - r_1) + \lambda s_2).$$

In the end, the two points end up at position $O(\lambda^{-1})$, and at distance $O(\lambda)$ from each other, which matches the description in Example 3.5(vi).

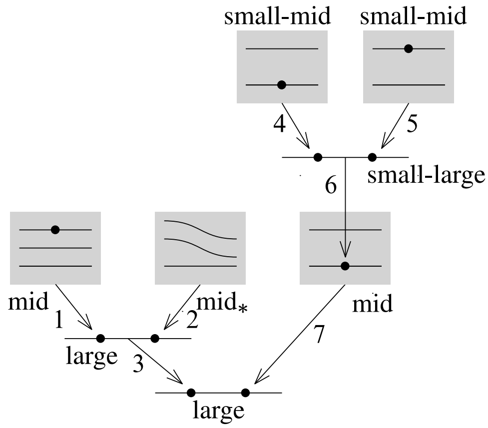

Figure 10: A generalized corner point; see Example 3.8.

Allowing some of the parameters to become zero yields a partial gluing process, which extends the chart obtained by gluing to include boundary points. In order for the relations (3-35) and (3-36) to make sense in this context, one multiplies them all by λ_e^{-1}, so as to get equations between monomials with nonnegative coefficients. One can think of this completely as a limit of the previous gluing process.

Example 3.8 Take the example from Figure 10. After preliminary simplifications, the relations between gluing parameters are $\lambda_1 = \lambda_2$, $\lambda_7 = \lambda_2\lambda_3$, $\lambda_4 = \lambda_5$ and, more importantly,

$$\lambda_2\lambda_3 = \lambda_5\lambda_6. \tag{3-41}$$

Hence, this point is not a classical corner in its moduli space. After gluing, the position of the two rightmost points is of order $\lambda_2\lambda_3$, and the distance between them is of order λ_6^{-1}. In the limit as all gluing parameters to go zero, $\lambda_2\lambda_3\lambda_6^{-1} = \lambda_5 \to 0$ by (3-41), as in the similar but simpler situation of Example 3.5(v).

It is convenient to pass from the multiplicative language of gluing parameters to the additive language of monoids. We define an abelian group G_T as follows. There is one generator g_e for each edge. For a vertex v, we define g_v to be the signed sum of g_e over a path from v_* to v, with signs according to orientations. The additive relations corresponding to the ones above are

$$g_v = 0 \quad \text{for a mid-scale vertex } v, \tag{3-42}$$

$$g_{v_+} + g_{v_-} = 0 \quad \text{if } v_+ \text{ is a turning point for } v_-. \tag{3-43}$$

Let $G_{T,\geq 0} \subset G_T$ be the submonoid generated by the g_e. The gluing parameters, including the degenerate cases where some are set to zero, are elements of $\mathrm{Hom}(G_T^{\geq 0}, \mathbb{R}^{\geq 0})$, where $\mathbb{R}^{\geq 0}$ is the multiplicative monoid.

Lemma 3.9 *G_T is a free abelian group, whose rank is the number of vertices of T which are neither mid-scale nor small-mid scale (in other words, the "other scales" in* (3-32)*).*

Proof Let E_T be the set of finite edges, and R_T be the set of relations. Our definition amounts to a short exact sequence

$$0 \to \mathbb{Z}^{R_T} \xrightarrow{\text{relations}} \mathbb{Z}^{E_T} \to G_T \to 0. \tag{3-44}$$

Any relation has a distinguished finite edge associated to it: for (3-42), the edge exiting v, and for (3-43), the edge exiting v_-. Those edges are pairwise different. Given an element of $\mathbb{Z}^{E_T}$, the coefficients for the distinguished edges give a splitting of the first map in (3-44), which implies freeness of the quotient. $\square$

Lemma 3.10 *$G_{T,\geq 0}$ is saturated, meaning that if $g \in G_T$ satisfies $mg \in G_{T,\geq 0}$ for some $m \geq 2$, then $g \in G_{T,\geq 0}$.*

Proof For this, it is simpler to work exclusively in terms of the g_v, and use (3-42) to drop the mid-scale vertices. Hence, let V_T be the set of all vertices which are not mid-scale. We start with $\mathbb{Z}^{V_T}$, and define G_T by quotienting out by (3-43). An element

$$\sum_{v \in V_T} m_v g_v \in \mathbb{Z}^{V_T} \tag{3-45}$$

is nonnegative if satisfies the following conditions. If v lies above v_* in our tree (meaning that the path from v to the root goes through v_*), then $m_v \leq 0$. If v_* lies above v, then $m_v \geq 0$. Finally, the m_v increase as one goes towards the root. As before, $G_{T,\geq 0}$ is the image of the nonnegative elements in the quotient G_T. Here is an equivalent form of the desired statement:

Claim *Given some element* (3-45)*, suppose that there are rational numbers $c_{v_-} \in [0,1]$, one for each small-mid-scale vertex v_-, such that*

$$\sum_{v \in V_T} m_v g_v + \sum_{v_- \text{small-mid}} c_{v_-}(g_{v_-} + g_{v_+}) \tag{3-46}$$

satisfies the nonnegativity condition. Then the same can be achieved with $c_{v_-} \in \{0,1\}$.

To prove this, we take (3-46) and then gradually modify the c_{v_-}. Take a turning point v_+. There can in principle be several corresponding small-mid scale vertices $v_{-,1}, \dots, v_{-,j}$. The coefficient of v_+ in (3-46) is then

$$m_{v_+} + c_{v_+}, \quad \text{where } c_{v_+} = \sum_{i=1}^{j} c_{v_{-,i}}. \tag{3-47}$$

If this is an integer, we do nothing. Otherwise, we increase (some of) the noninteger $c_{v_{-,i}}$ until the resulting expression (3-47) becomes equal to the next larger integer. Let's apply this to all turning points. The outcome is that now we have an expression (3-46) which still satisfies the nonnegativity condition, and where the coefficients of all turning points are integers. In a second pass, we change the coefficients of small-mid scale vertices again, but without affecting (3-47), to make all of them integers. The situation is, simplifying the notation, that we have noninteger $c_1, \dots, c_k \in [0, 1]$ such that $c_1 + \dots + c_k$ is an integer; and we then need to change them to be either 0 or 1, while preserving the sum, something that's clearly possible. Having done that, we have justified our claim. □

Lemma 3.11 *$G_{T,\geq 0}$ is sharp, meaning it contains no nontrivial pair of elements $\pm g$.*

Proof We know that $\mathrm{Hom}(G_{T,\geq 0}, \mathbb{R}^{\geq 0})$ recovers the space of gluing parameters, including degenerate ones. In particular, there is a distinguished point where all gluing parameters are set to zero, which is the zero map. Composing that with a homomorphism $\mathbb{Z} \to G_{T,\geq 0}$ would mean that the zero map $\mathbb{Z} \to \mathbb{R}^{\geq 0}$ is a group homomorphism, which is nonsense. □

Lemmas 3.9–3.11 say that $G_T^{\geq 0}$ is a toric monoid (terminology as in [36]). For the space of gluing parameters, this is precisely what defines a generalized corner.

4 The formal group structure

This section carries out versions of our main constructions in an idealized context, where the technicalities of symplectic topology have been replaced by a general operadic framework (this degree of abstraction comes with its own occasional complications). The primary objects under consideration will be chain complexes which are algebras over the Fulton–MacPherson operad. Abstractly speaking, in view of [53, Theorem 1.1], this

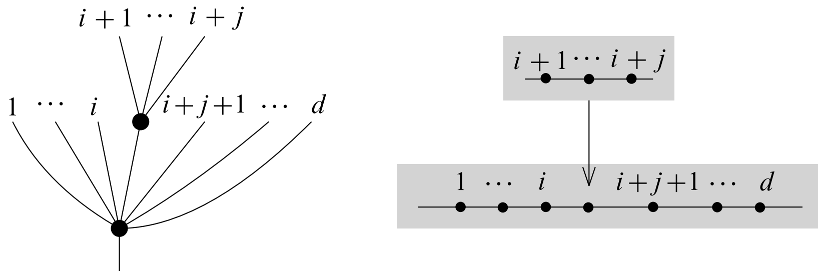

Figure 11: The boundary faces of S_d appearing in (4-1), as trees and as strata of the compactified configuration space.

situation is not more general than the purely algebraic one mentioned in Remark 2.12. However, that viewpoint lacks the geometric explicitness which is useful for applications to symplectic topology.

4a Associahedra

Consider the singular chain complexes of the associahedra, $C_*(S_d)$. These inherit the structure of a nonsymmetric operad, using the maps induced by (3-2) as well as the shuffle (Eilenberg–Mac Lane or Eilenberg–Zilber) product. One can inductively construct "fundamental chains" $[S_d] \in C_{d-2}(S_d)$ such that $[S_2] = [\text{point}]$ and

$$\partial[S_d] = \sum_{ij} (-1)^{(d-i-j)j+i} T_{ij,*}([S_{d-j+1}] \times [S_j]). \tag{4-1}$$

Here, the sum is over pairs (i, j) corresponding to trees T_{ij} with two vertices, of valences $j+1$ and $d-j+2$, respectively; and where the unique finite edge is the $(i+1)^{\text{st}}$ incoming edge of the first vertex ($0 \leq i < d-j+1$); see Figure 11. We take the shuffle product (here just denoted by $\times$) of the fundamental chains $[S_{d-j+1}]$ and $[S_j]$, and then map that to $C_*(S_d)$ by the chain-level map induced by (3-2), denoted here by $T_{ij,*}$. The sign takes into account the co-orientations of the boundary faces. In view of (4-1), $[S_d]$ has a preferred lift to a cycle for the pair $(S_d, \partial S_d)$, whose homology class is then a fundamental class in the standard sense, compatible with the orientations described in Section 4a.

Our standing convention is that chain complexes are cohomologically graded, hence we now switch to the grading-reversed version $C_{-*}(S_d)$. By an algebra over the chain-level Stasheff operad, we mean a chain complex of free abelian groups $\mathcal{A}$, which comes with maps

$$C_{-*}(S_d) \otimes \overbrace{\mathcal{A} \otimes \cdots \otimes \mathcal{A}}^{d} \to \mathcal{A}, \tag{4-2}$$

compatible with the composition maps induced by (3-2). Let's evaluate these maps at $[S_d] \otimes a_1 \otimes \cdots \otimes a_d$, multiply with a sign $(-1)^*$, where

$$* = (d-1)|a_1| + (d-2)|a_2| + \cdots + |a_{d-1}|, \tag{4-3}$$

and denote the outcome by $\mu^d_{\mathcal{A}}(a_1, \ldots, a_d)$. These maps, together with $\mu^1_{\mathcal{A}} = -d_{\mathcal{A}}$, make $\mathcal{A}$ into an A_∞–ring. The associativity equations (2-1) are a direct consequence of (4-1). Homological unitality is not part of this framework, hence has to be imposed as a separate property.

Remark 4.1 It is maybe appropriate to recall briefly how the signs work out. If we denote the operation (4-2) by $o^d_{\mathcal{A}}$, the starting point is its chain map property, which together with (4-1) yields

$$\begin{aligned}\sum_{ij} (-1)^{(d+1)j+i(j+1)} o^d_{\mathcal{A}}(T_{ij,*}([S_{d-j+1}] \times [S_j]) \otimes a_1 \otimes \cdots \otimes a_d)& \\ + \text{(terms involving } d_{\mathcal{A}}) = 0.&\end{aligned} \tag{4-4}$$

The operad property, not forgetting the Koszul signs, transforms this into

$$\begin{aligned}\sum_{ij} (-1)^{(d+1)j+i+j\maltese_i} o^{d-j+1}_{\mathcal{A}}\big([S_{d-j+1}] \otimes a_1 \otimes \cdots& \\ \otimes o^j_{\mathcal{A}}([S_j] \otimes a_{i+1} \otimes \cdots \otimes a_{i+j}) \otimes \cdots \otimes a_d\big)& \\ + \text{(terms involving } d_{\mathcal{A}}) = 0;&\end{aligned} \tag{4-5}$$

or, in terms of the A_∞–operations,

$$\begin{aligned}(-1)^* \sum_{ij} (-1)^{\maltese_i} \mu^{d-j+1}_{\mathcal{A}}(a_1, \ldots, \mu^j_{\mathcal{A}}(a_{i+1}, \ldots, a_{i+j}), \ldots, a_d)& \\ + \text{(terms involving } d_{\mathcal{A}}) = 0,&\end{aligned} \tag{4-6}$$

with $*$ as in (4-3). The sum in (4-6) is over $2 \leq j \leq d-1$, but only because we have omitted the differential terms, which are

$$\begin{aligned}\sum_i (-1)^{d+\maltese_i+i} o_{\mathcal{A}}([S_d] \otimes a_1 \otimes \cdots \otimes d_{\mathcal{A}} a_{i+1} \otimes \cdots \otimes a_d)& \\ - d_{\mathcal{A}}(o^d_{\mathcal{A}}([S_d] \otimes a_1 \otimes \cdots a_d))& \\ = (-1)^* \sum_i (-1)^{\maltese_i} \mu^d_{\mathcal{A}}(a_1, \ldots, \mu^1_{\mathcal{A}}(a_{i+1}), \ldots, a_d)& \\ + (-1)^* \mu^1_{\mathcal{A}}(\mu^d_{\mathcal{A}}(a_1, \ldots, a_d)).&\end{aligned} \tag{4-7}$$

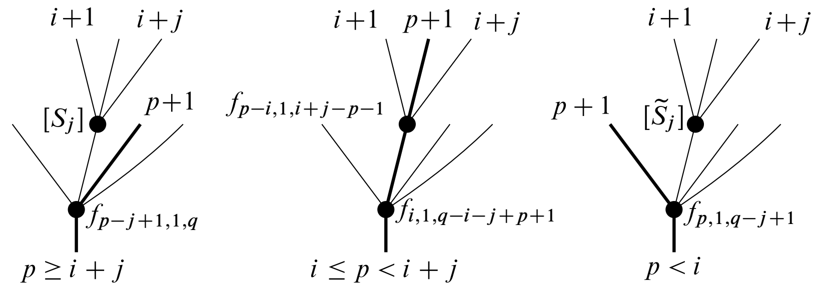

Figure 12: Graphical representation of the terms in (4-8).

4b Dependence on the fundamental chains

Suppose we are given two sequences of fundamental chains $[S_d]$ and $[\widetilde{S}_d]$, each of which separately satisfies (4-1). To relate them, we want to make further choices of fundamental chains, which have a mixed boundary property:

$$
\begin{aligned}
&f_{p,1,q} \in C_{d-2}(S_d), \quad \text{where } p,q \geq 0 \text{ and } d = p+1+q,\\
&f_{p,1,0} = [S_{p+1}],\\
&f_{0,1,q} = [\widetilde{S}_{q+1}],\\
&\partial f_{p,1,q} =\\
&\sum_{ij} (-1)^{(d-i-j)j+i} T_{ij,*} \begin{cases} f_{p-j+1,1,q} \times [S_j] & \text{if } p \geq i+j,\\ f_{i,1,p+q+1-i-j} \times f_{p-i,1,i+j-p-1} & \text{if } i \leq p < i+j,\\ f_{p,1,q-j+1} \times [\widetilde{S}_j] & \text{if } p < i. \end{cases}
\end{aligned}
\tag{4-8}
$$

Graphically, one can think of (4-8) as follows. Let's mark the $(p+1)^{\text{st}}$ leaf of our planar trees. Vertices that lie on the unique path connecting that leaf to the root correspond to factors carrying an appropriate f chain, while the remaining ones always carry $[S]$ or $[\widetilde{S}]$ chains, depending on whether they lie to the left or right of the path; see Figure 12.

Let $\mu_{\mathcal{A}}$ and $\widetilde{\mu}_{\mathcal{A}}$ be the A_∞–ring structures associated to $[S_d]$ and $[\widetilde{S}_d]$. In the same way, the action of $f_{p,1,q}$ gives rise to operations

$$
\begin{gathered}
\phi_{\mathcal{A}}^{p,1,q} : \mathcal{A}^{\otimes p+q+1} \to \mathcal{A}[1-p-q], \quad \text{where } p+1+q \geq 2,\\
\phi_{\mathcal{A}}^{p,1,0} = \mu_{\mathcal{A}}^{p+1}, \quad \phi_{\mathcal{A}}^{0,1,q} = \widetilde{\mu}_{\mathcal{A}}^{q+1},
\end{gathered}
\tag{4-9}
$$

which, as before, we extend by setting

$$\phi_{\mathcal{A}}^{0,1,0} = -d_{\mathcal{A}}.$$

The relations inherited from (4-8) are

$$
\begin{aligned}
\text{(4-10)}\quad & \sum_{p\geq i+j} (-1)^{\maltese_i} \phi_{\mathcal{A}}^{p-j+1,1,q}\big(a_1,\dots,\mu_{\mathcal{A}}^{j}(a_{i+1},\dots,a_{i+j}),\dots,a_p;a_{p+1};\\
&\qquad\qquad a_{p+2},\dots,a_{p+q+1}\big)\\
&+\sum_{i\leq p<i+j} (-1)^{\maltese_i} \phi_{\mathcal{A}}^{i,1,p+q+1-i-j}\big(a_1,\dots;\\
&\qquad \phi_{\mathcal{A}}^{p-i,1,i+j-p-1}(a_{i+1},\dots,a_p;a_{p+1};a_{p+2},\dots,a_{i+j});\\
&\qquad\qquad \dots,a_{p+q+1}\big)\\
&+\sum_{p<i} (-1)^{\maltese_i} \phi_{\mathcal{A}}^{p,1,q-j+1}\big(a_1,\dots,a_p;a_{p+1};\\
&\qquad a_{p+2},\dots,\tilde{\mu}_{\mathcal{A}}^{j}(a_{i+1},\dots,a_{i+j}),\dots,a_{p+q+1}\big)\\
&=0.
\end{aligned}
$$

Remark 4.2 The operations (4-9) equip the shifted space $\mathcal{A}[1]$ with the structure of an A_∞–bimodule, where μ acts on the left and $\tilde{\mu}$ on the right; see eg [64, equation (2.5)]; the shift is there to match sign conventions.

In a second step, we find fundamental chains

$$
\begin{aligned}
\text{(4-11)}\quad & g_{p,q} \in C_{d-1}([0,1]\times S_d), \quad \text{where } p,q>0 \text{ and } d=p+q,\\
& \partial g_{p,q} = \{1\}\times f_{p-1,1,q} - \{0\}\times f_{p,1,q-1}\\
&\qquad + \sum_{ij} (-1)^{(d-i-j)j+i}\\
&\qquad\qquad \cdot T_{ij,*}\begin{cases} -g_{p-j+1,q}\times[S_j] & \text{if } p\geq i+j,\\ (-1)^{d-j} f_{i,1,q-i+p-j}\times g_{p-i,i+j-p} & \text{if } i\leq p<i+j,\\ -g_{p,q-j+1}\times[\tilde{S}_j] & \text{if } p<i.\end{cases}
\end{aligned}
$$

When compared to (4-1) and (4-8), the spaces involved have acquired an additional $[0,1]$ factor; hence, we should really write $\mathrm{id}_{[0,1]}\times T_{ij,*}$. The graphical representation involves drawing a dividing line between the first p and last q leaves of our trees. In the first two summands in (4-11), we remove that dividing line and instead mark the leaves that are on either side of it, leading to the appearance of two f terms. For the remaining summands, vertices to the left (resp. right) of the dividing line carry $[S]$ (resp. $[\tilde{S}]$) chains; see Figure 13. If the dividing line ends at the top vertex, which is the middle case in both (4-11) and Figure 13, the finite edge of the tree becomes the marked edge of the bottom vertex, which explains how that vertex carries an f term.

Let's take the image of (4-11) under projection to S_d. Its action under the operad structure, with additional signs inserted as in (4-3), gives operations

$$
\text{(4-12)}\qquad \psi_{\mathcal{A}}^{p,q}\colon \mathcal{A}^{\otimes p+q}\to\mathcal{A}[1-p-q], \quad \text{where } p,q>0,
$$

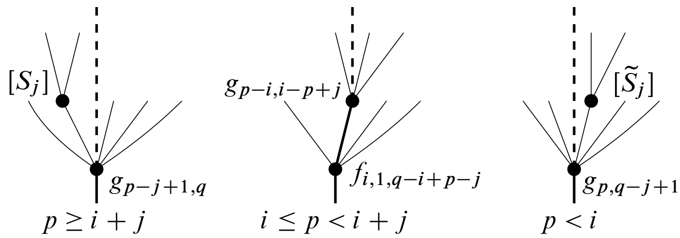

Figure 13: Graphical representation of the $\sum_{ij}$ in (4-11).

which we complement by setting $\psi_{\mathcal{A}}^{1,0} = -\mathrm{id}_{\mathcal{A}}$ and $\psi_{\mathcal{A}}^{0,1} = \mathrm{id}_{\mathcal{A}}$. These satisfy

$$\begin{aligned}
\text{(4-13)} \quad & \sum_{p \geq i+j} (-1)^{\maltese_i} \psi_{\mathcal{A}}^{p-j+1,q}\big(a_1,\dots,\mu_{\mathcal{A}}^j(a_{i+1},\dots,a_{i+j}),\dots,a_p; \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad a_{p+1},\dots,a_{p+q}\big) \\
& - \sum_{i<p<i+j} \phi_{\mathcal{A}}^{i,1,p+q-i+j}\big(a_1,\dots,a_i; \\
&\qquad\qquad \psi_{\mathcal{A}}^{p-i,i+j-p}(a_{i+1},\dots,a_p;a_{p+1},\dots,a_{i+j});a_{i+j+1},\dots,a_{p+q}\big) \\
& + \sum_{p\leq i} (-1)^{\maltese_i} \psi_{\mathcal{A}}^{p,q-j+1}\big(a_1,\dots,a_p; \\
&\qquad\qquad\qquad a_{p+1},\dots,\tilde{\mu}_{\mathcal{A}}^j(a_{i+1},\dots,a_{i+j}),\dots,a_{p+q}\big) \\
&= 0.
\end{aligned}$$

Note that (4-13) contains terms which correspond to the boundary faces $\{0\} \times S_d$ and $\{1\} \times S_d$:

$$\begin{aligned}
\text{(4-14)} \quad & -\phi_{\mathcal{A}}^{p-1,1,q}(a_1,\dots,a_{p-1};\psi_{\mathcal{A}}^{1,0}(a_p);a_{p+1},\dots,a_{p+q}) \\
&\qquad\qquad - \phi_{\mathcal{A}}^{p,1,q-1}(a_1,\dots,a_p;\psi_{\mathcal{A}}^{0,1}(a_{p+1});a_{p+2},\dots,a_{p+q}) \\
&= \phi_{\mathcal{A}}^{p-1,1,q}(a_1,\dots,a_{p-1};a_p;a_{p+1},\dots,a_{p+q}) \\
&\qquad\qquad - \phi_{\mathcal{A}}^{p,1,q-1}(a_1,\dots,a_p;a_{p+1};a_{p+2},\dots,a_{p+q}).
\end{aligned}$$

Example 4.3 The simplest instance of (4-13), bearing in mind the conventions for $\phi^{0,1,0}$, $\phi^{1,1,0}$ and $\phi^{0,1,1}$, is

$$\begin{aligned}
\text{(4-15)} \quad & \psi_{\mathcal{A}}^{1,1}(\mu_{\mathcal{A}}^1(a_1);a_2) + (-1)^{\|a_1\|}\psi_{\mathcal{A}}^{1,1}(a_1;\mu_{\mathcal{A}}^1(a_2)) \\
&\qquad = \mu_{\mathcal{A}}^1(\psi_{\mathcal{A}}^{1,1}(a_1;a_2)) + \mu_{\mathcal{A}}^2(a_1,a_2) - \tilde{\mu}_{\mathcal{A}}^2(a_1,a_2).
\end{aligned}$$

This says that $(-1)^{|a_1|}\psi^{1,1}(a_1,a_2)$ is a chain homotopy relating the two versions of multiplication.

Remark 4.4 Following up on our last observation, one can give the following interpretation of (4-13). Recall from Remark 4.2 that the operations ϕ equip $\mathcal{A}$ (here, we undo the shift for simplicity) with an A_∞–bimodule structure. By construction, this is isomorphic to $(\mathcal{A},\mu_{\mathcal{A}})$ as a left module over itself, and to $(\mathcal{A},\tilde{\mu}_{\mathcal{A}})$ as a right module. Correspondingly, one has two bimodule maps

$$(\mathcal{A},\mu_{\mathcal{A}})\otimes_{\mathbb{Z}}(\mathcal{A},\tilde{\mu}_{\mathcal{A}}) \underset{\tilde{\rho}_{\mathcal{A}}}{\overset{\rho_{\mathcal{A}}}{\rightrightarrows}} (\mathcal{A},\phi_{\mathcal{A}}), \tag{4-16}$$

given by

$$\begin{aligned} &\rho_{\mathcal{A}}^{p-1,1,q-1}(a_1,\dots;a_p\otimes a_{p+1};\dots,a_{p+q}) \\ &\qquad\qquad\qquad\qquad = \pm\phi_{\mathcal{A}}^{p,1,q-1}(a_1,\dots;a_{p+1};\dots,a_{p+q}), \\ &\tilde{\rho}_{\mathcal{A}}^{p-1,1,q-1}(a_1,\dots;a_p\otimes a_{p+1};\dots,a_{p+q}) \\ &\qquad\qquad\qquad\qquad = \pm\phi_{\mathcal{A}}^{p-1,1,q}(a_1,\dots;a_p;\dots,a_{p+q}). \end{aligned} \tag{4-17}$$

In these terms, (4-13) says that ψ provides a homotopy between ρ and $\tilde{\rho}$.

It is worth noting that homological unitality, when it holds, can be used to simplify the picture. Namely, suppose that $\mu_{\mathcal{A}}$ and $\tilde{\mu}_{\mathcal{A}}$ are both homologically unital, with a priori different units $e_{\mathcal{A}}$ and $\tilde{e}_{\mathcal{A}}$. Then a bimodule map as in (4-16) is determined up to homotopy by the image of $[e_{\mathcal{A}}\otimes\tilde{e}_{\mathcal{A}}]$ in $H^0(\mathcal{A})$. In our situation, the two classes are

$$[\mu^2_{\mathcal{A}}(e_{\mathcal{A}},\tilde{e}_{\mathcal{A}})]=[\tilde{e}_{\mathcal{A}}] \quad\text{and}\quad [\tilde{\mu}^2_{\mathcal{A}}(e_{\mathcal{A}},\tilde{e}_{\mathcal{A}})]=[e_{\mathcal{A}}], \tag{4-18}$$

so the existence of a homotopy ψ just amounts to saying that the two units are, after all, cohomologous. Similarly, the different choices of ψ form an affine space over $H^{-1}(\mathcal{A})$.

Let's define an A_∞–ring structure on

$$\mathcal{H}=\mathcal{A}\otimes\mathcal{I}=\mathcal{A}u\oplus\mathcal{A}\tilde{u}\oplus\mathcal{A}v, \tag{4-19}$$

where $\mathcal{I}$ is the noncommutative interval (2-3), as follows. The differential $\mu^1_{\mathcal{H}}$ is as in (2-4). The nonzero higher A_∞–operations, for $d\geq 2$ and $p,q>0$, are

$$\begin{aligned} &\mu^d_{\mathcal{H}}(a_1\otimes u,\dots,a_d\otimes u)=\mu^d_{\mathcal{A}}(a_1,\dots,a_d)\otimes u, \\ &\mu^{p+q}_{\mathcal{H}}(a_1\otimes u,\dots,a_p\otimes u,a_{p+1}\otimes\tilde{u},\dots,a_{p+q}\otimes\tilde{u}) \\ &\qquad\qquad\qquad = (-1)^{\maltese_{p+q}}\psi^{p,q}_{\mathcal{A}}(a_1,\dots,a_p;a_{p+1},\dots,a_{p+q})\otimes v, \\ &\mu^d_{\mathcal{H}}(a_1\otimes\tilde{u},\dots,a_d\otimes\tilde{u})=\tilde{\mu}^d_{\mathcal{A}}(a_1,\dots,a_d)\otimes\tilde{u}, \\ &\mu^{p+q+1}_{\mathcal{H}}(a_1\otimes u,\dots,a_p\otimes u,a_{p+1}\otimes v,a_{p+2}\otimes\tilde{u},\dots,a_{p+q+1}\otimes\tilde{u}) \\ &\qquad = (-1)^{\maltese_{p+q+1}-\maltese_{p+1}}\phi^{p,1,q}(a_1,\dots,a_p;a_{p+1};a_{p+2},\dots,a_{p+q+1})\otimes v. \end{aligned} \tag{4-20}$$

(This generalizes the previous (2-4), which corresponds to the diagonal A_∞–bimodule structure and vanishing ψ.) The A_∞–associativity relations follow directly from (4-10) and (4-13).

Remark 4.5 As a check on the signs, consider the associativity relation for

$$(a_1 \otimes u, \dots, a_p \otimes u, a_{p+1} \otimes \tilde{u}, \dots, a_{p+q} \otimes \tilde{u}),$$

and more specifically the v–component of that relation. This turns out to be exactly (4-13) multiplied by $(-1)^{\maltese_{p+q}+1}$. The crucial terms, compare (4-14), are

$$\begin{aligned}(-1)^{\maltese_{p-1}} \mu_{\mathcal{H}}^{p+q}(a_1 \otimes u, \dots, v\text{–component of } \mu^1_{\mathcal{H}}(a_p \otimes u), a_{p+1} \otimes \tilde{u}, \dots) \\ = (-1)^{\maltese_p+1} \mu_{\mathcal{H}}^{p+q}(a_1 \otimes u, \dots, a_p \otimes v, a_{p+1} \otimes \tilde{u}, \dots) \\ = (-1)^{\maltese_{p+q}+1} \phi_{\mathcal{A}}^{p-1,1,q}(a_1, \dots, a_{p-1}; a_p; a_{p+1}, \dots, a_{p+q}) \otimes v\end{aligned} \tag{4-21}$$

and

$$\begin{aligned}(-1)^{\maltese_p} \mu_{\mathcal{H}}^{p+q}(a_1 \otimes u, \dots, a_p \otimes u, v\text{–component of } \mu^1_{\mathcal{H}}(a_{p+1} \otimes \tilde{u}), \dots) \\ = (-1)^{\maltese_{p+1}} \mu_{\mathcal{H}}^{p+q}(a_1 \otimes u, \dots, a_p \otimes u, a_{p+1} \otimes v, \dots) \\ = (-1)^{\maltese_{p+q}} \phi_{\mathcal{A}}^{p,1,q-1}(a_1, \dots, a_p; a_{p+1}; a_{p+2}, \dots, a_{p+q}) \otimes v.\end{aligned} \tag{4-22}$$

By construction, the projections (2-5) are A_∞–homomorphisms from $\mu_{\mathcal{H}}$ to $\mu_{\mathcal{A}}$ and $\tilde{\mu}_{\mathcal{A}}$, respectively, and also chain homotopy equivalences. By taking a homotopy inverse (Lemma 2.1) of one projection, and composing with the other projection, we get an A_∞–homomorphism

$$(\mathcal{A}, \mu_{\mathcal{A}}) \to (\mathcal{A}, \tilde{\mu}_{\mathcal{A}}), \tag{4-23}$$

whose linear part is homotopic to the identity (one can achieve that it is exactly the identity). For a completely satisfactory statement, one would need to prove that (4-23) is itself independent of the choice of (4-9) and (4-12) up to homotopy of A_∞–homomorphisms; and also, that the composition of two maps (4-23) is again a map of the same type, up to homotopy. This would use higher analogues of $\mathcal{I}$. For the sake of brevity, we will not carry it out here.

4c Fulton–MacPherson spaces and colored multiplihedra

One defines the structure of an algebra over $C_{-*}(\mathrm{FM}_d)$ on a chain complex $\mathcal{C}$ by maps analogous to (4-2), with the additional stipulation of Sym_d–invariance. On the cohomology level, $H^*(\mathcal{C})$ becomes a Gerstenhaber algebra. The chain-level structure is a classical topic in algebraic topology (E_2–algebras; see eg [51; 14; 53; 68]). For our purpose, only part of that structure is relevant — that part, maybe surprisingly, does not include the fundamental chains $[\mathrm{FM}_d] \in C_{2d-3}(\mathrm{FM}_d)$ and the resulting L_∞–structure; in fact, the chains relevant for us have dimension $\leq d$.

First of all, having chosen fundamental chains $[S_d]$ for the Stasheff associahedra, one can map them to $C_{d-2}(\mathrm{FM}_d)$ via (3-5), and their action turns $\mathcal{C}$ into an A_∞–ring. As before, one has to require homological unitality separately. Next, choose fundamental chains $[\mathrm{MWW}_{d_1,\dots,d_r}] \in C_{d-1}(\mathrm{MWW}_{d_1,\dots,d_r})$ for the colored multiplihedra, which satisfy the analogue of (4-1). It is worthwhile writing this down:

$$\begin{aligned}\partial[\mathrm{MWW}_{d_1,\dots,d_r}] = {} & \\ \sum_{ijk} (-1)^{(d_k-i-j+d_{k+1}+\cdots+d_r)j+(d_1+\cdots+d_{k-1}+i+1)} & \\ \cdot\, T_{ijk,*}([\mathrm{MWW}_{d_1,\dots,d_k-j+1,\dots,d_r}]\times[S_j]) & \\ + \sum_{\text{partitions}} (-1)^{\diamond} T_{d_{1,1},\dots,d_{j,r},*}([S_j]\times[\mathrm{MWW}_{d_{1,1},\dots,d_{1,r}}]\times\cdots\times[\mathrm{MWW}_{d_{j,1},\dots,d_{j,r}}]). &\end{aligned} \tag{4-24}$$

The second sum is over all $j \geq 2$ and partitions

$$d_1 = d_{1,1} + \cdots + d_{j,1},\ \dots,\ d_r = d_{1,r} + \cdots + d_{j,r}$$

such that $d_{i,1} + \cdots + d_{i,r} > 0$ for each $i = 1, \dots, j$. The sign there is given by

$$\begin{aligned}\diamond = \sum_{i_1<i_2;\ k_1>k_2} d_{i_1,k_1} d_{i_2,k_2} + (j-1)(d_{1,1} + \cdots + d_{1,r} - 1) & \\ + (j-2)(d_{2,1} + \cdots + d_{2,r} - 1) + \cdots . &\end{aligned} \tag{4-25}$$

Example 4.6 Examples of the degenerate configurations corresponding to the terms in (4-24) are shown in Figure 14. (The trees T_{ijk} and $T_{d_{1,1},\dots,d_{j,r}}$ can be inferred from looking at those, so we will not define them explicitly.) Figures 3 and 5 illustrate the orientation issues: in both of them, the actual moduli space has the standard orientation of the plane, and the arrows show the orientations of the boundary strata arising from (3-11).

Choose maps (3-13), take the images of the fundamental chains under those maps, and let them act on $\mathcal{C}$. The outcomes are operations

$$\beta_{\mathcal{C}}^{d_1,\dots,d_r} : \mathcal{C}^{\otimes d} \to \mathcal{C}[1-d]. \tag{4-26}$$

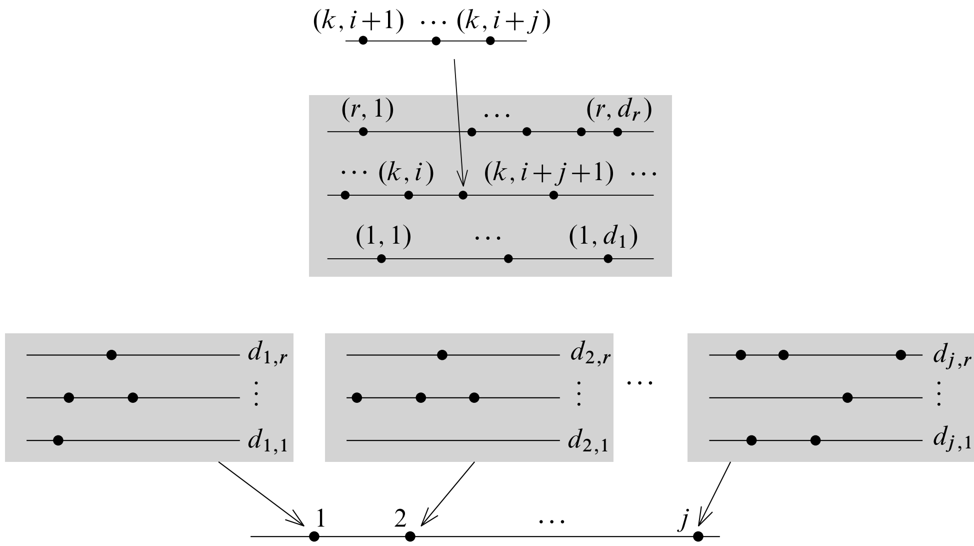

Figure 14: Top: a summand in the first sum in (4-24). Bottom: a summand in the second sum in (4-24).

In their definition, we insert signs as in (4-3); for

$$\beta_{\mathcal{C}}^{d_1,\dots,d_r}(c_{1,1},\dots,c_{1,d_1};\dots;c_{r,1},\dots,c_{r,d_r})$$

this means $(-1)^*$ with

$$* = (d_1+\dots+d_r-1)|c_{1,1}| + (d_1+\dots+d_r-2)|c_{1,2}| + \cdots + (d_1+\dots+d_{r-1})|c_{1,d_1}| + (d_1+\dots+d_{r-1}-1)|c_{2,1}| + \cdots . \tag{4-27}$$

For $\mathrm{MWW}_{0,\dots,0,1,0,\dots,0} = \text{point}$, where there is no corresponding Fulton–MacPherson space, we artificially set

$$\beta_{\mathcal{C}}^{0,\dots,0,1,0,\dots,0} = \mathrm{id}_{\mathcal{C}}. \tag{4-28}$$

As a consequence of (4-24),

$$\begin{aligned}\sum_{ijk} (-1)^{\maltese_{k,i}} \beta_{\mathcal{C}}^{d_1,\dots,d_k-j+1,\dots,d_r}\big(&c_{1,1},\dots,c_{1,d_1};\dots;\\ &c_{k,1},\dots,\mu_{\mathcal{C}}^j(c_{k,i+1},\dots,c_{k,i+j}),\dots,c_{k,d_k};\dots;c_{r,1},\dots,c_{r,d_r}\big)\\ = \sum_{\text{partitions}} (-1)^{\heartsuit} \mu_{\mathcal{C}}^j\big(&\beta_{\mathcal{C}}^{d_{1,1},\dots,d_{1,r}}(c_{1,1},\dots,c_{1,d_{1,1}};\dots;c_{r,1},\dots,c_{r,d_{1,r}}),\dots,\\ &\beta_{\mathcal{C}}^{d_{j,1},\dots,d_{j,r}}(c_{1,d_1-d_{j,1}+1},\dots,c_{1,d_1};\dots;c_{d_r-d_{j,r}+1},\dots,c_{r,d_r})\big).\end{aligned} \tag{4-29}$$

Here, the sums are over indexing sets as in (4-24), except that we now additionally allow the differential $\mu_{\mathcal{C}}^1 = d_{\mathcal{C}}$. Recall that by construction, the map (3-13) forgets factors of $\mathrm{MWW}_{0,\dots,0,1,0,\dots,0} = \text{point}$. Algebraically, this corresponds to the places where (4-28) appears in (4-29). The ✠ symbol is the sum of reduced degrees of all c which precede the μ; and $\heartsuit$ yields the Koszul sign that corresponds to permuting the $c_{k,i}$ from their original order into the order in which they appear on the right-hand side of (4-29), but using reduced degrees $\|c_{k,i}\|$.

Remark 4.7 The operations (4-26) constitute an A_∞–multihomomorphism with r entries $\mathcal{C} \times \cdots \times \mathcal{C} \to \mathcal{C}$ (the single-object version of an A_∞–multifunctor; see [4, Definition 8.8], or closer to our context, the discussion of the $r = 2$ case in [9, Section 4.5]). What we will study later on amounts to the action of those A_∞–multihomomorphisms on Maurer–Cartan elements. One can argue that the homomorphisms themselves should be the center of attention;[3] in the interest of keeping the discussion concrete, we have chosen not to take that route.

Example 4.8 In view of (4-28) and (4-29), $\beta^2_{\mathcal{C}}$ satisfies

$$\begin{aligned}\mu^1_{\mathcal{C}}(\beta^2_{\mathcal{C}}(c_1,c_2)) - \beta^2_{\mathcal{C}}(\mu^1_{\mathcal{C}}(c_1),c_2) - (-1)^{\|c_1\|}\beta^2_{\mathcal{C}}(c_1,\mu^1_{\mathcal{C}}(c_2)) \\ = \beta^1_{\mathcal{C}}(\mu^2_{\mathcal{C}}(c_1,c_2)) - \mu^2_{\mathcal{C}}(\beta^1_{\mathcal{C}}(c_1),\beta^1_{\mathcal{C}}(c_2)) = 0,\end{aligned} \tag{4-30}$$

which means that $(-1)^{|c_1|}\beta^2_{\mathcal{C}}(c_1,c_2)$ is a chain map of degree -1. Geometrically, the reason is that the image of the fundamental chain under $\mathrm{MWW}_2 \to \mathrm{FM}_2$ is a one-cycle. However, this cycle is supported at a single point of $\mathrm{FM}_2 \cong S^1$, hence is necessarily nullhomologous. This implies that $\beta^2_{\mathcal{C}}$ is chain homotopic to zero.

Example 4.9 The first substantially nontrivial case is $\beta^{1,1}_{\mathcal{C}}$, which satisfies

$$\begin{aligned}\mu^1_{\mathcal{C}}\beta^{1,1}_{\mathcal{C}}(c_1;c_2) - \beta^{1,1}_{\mathcal{C}}(\mu^1_{\mathcal{C}}(c_1);c_2) - (-1)^{\|c_1\|}\beta^{1,1}_{\mathcal{C}}(c_1;\mu^1_{\mathcal{C}}(c_2)) \\ = -\mu^2_{\mathcal{C}}(c_1,c_2) - (-1)^{\|c_1\|\cdot\|c_2\|}\mu^2_{\mathcal{C}}(c_2,c_1).\end{aligned} \tag{4-31}$$

In more conventional terminology, $(-1)^{|c_1|}\beta^{1,1}_{\mathcal{C}}(c_1;c_2)$ is the $\circ$ operation which shows homotopy commutativity of the product on $H^*(\mathcal{C})$.

Definition 4.10 Fix an adic ring N (Definition 1.1). Given $\gamma_1,\dots,\gamma_r \in \mathcal{C}^1 \mathbin{\widehat{\otimes}} N$, define

$$\Pi^r_{\mathcal{C}}(\gamma_1,\dots,\gamma_r) = \sum_{\substack{d_1,\dots,d_r \geq 0,\\ d_1+\dots+d_r>0}} \beta^{d_1,\dots,d_r}_{\mathcal{C}}(\overbrace{\gamma_1,\dots,\gamma_1}^{d_1};\dots;\overbrace{\gamma_r,\dots,\gamma_r}^{d_r}). \tag{4-32}$$

[3] Meaning that Proposition 4.20 should be understood as a consequence of a composition property of the A_∞–multihomomorphisms up to homotopy; and similarly that Corollary 4.25 should be true because for $r = 1$, one gets an A_∞–endomorphism of $\mathcal{C}$ which is homotopy equivalent to the identity.

Suppose that we have $(\gamma_1, \dots, \gamma_r)$ as well as, for some k, another element $\tilde{\gamma}_k$.[4] Then, define a linear endomorphism of $\mathcal{C} \mathbin{\hat{\otimes}} N$ by a generalization of (2-11),

$$P^{r,k}_{\mathcal{C}}(\xi) = \sum_{d_1,\dots,d_r} \sum_{p+q+1=d_r} \beta_{\mathcal{C}}^{d_1,\dots,d_k,\dots,d_r}\bigl(\overbrace{\gamma_1,\dots,\gamma_1}^{d_1};\dots;\overbrace{\gamma_k,\dots,\gamma_k}^{p},\xi, \underbrace{\tilde{\gamma}_k,\dots,\tilde{\gamma}_k}_{q};\dots;\underbrace{\gamma_r,\dots,\gamma_r}_{d_r}\bigr). \tag{4-33}$$

The definitions, taking (4-28) into account, have the following immediate consequences:

$$\Pi^r_{\mathcal{C}}(\gamma_1,\dots,\gamma_r) = \gamma_1 + \dots + \gamma_r \mod N^2, \tag{4-34}$$

$$\Pi^r_{\mathcal{C}}(\gamma_1,\dots,\gamma_k+\xi,\dots,\gamma_r) = \Pi_{\mathcal{C}}(x) + P^{r,k}_{\mathcal{C}}(\xi) + (\text{order} \geq 2 \text{ terms in } \xi), \tag{4-35}$$

$$\xi \in \mathcal{C} \mathbin{\hat{\otimes}} N^m \implies P^{r,k}_{\mathcal{C}}(\xi) = \xi \mod N^{m+1}, \tag{4-36}$$

$$\Pi^r_{\mathcal{C}}(\gamma_1,\dots,\gamma_r) - \Pi^r_{\mathcal{C}}(\gamma_1,\dots,\tilde{\gamma}_k,\dots,\gamma_r) = P^{r,k}_{\mathcal{C}}(\gamma_k - \tilde{\gamma}_k). \tag{4-37}$$

In (4-35), the endomorphism $P^{r,k}_{\mathcal{C}}$ is with respect to $\tilde{\gamma}_k = \gamma_k$. The two subsequent equations, in contrast, use a general $\tilde{\gamma}_k$.

Lemma 4.11 *If $\gamma_1,\dots,\gamma_r$ are Maurer–Cartan elements* (1-1), *then so is*

$$\gamma = \Pi^r_{\mathcal{C}}(\gamma_1,\dots,\gamma_r).$$

Moreover, the equivalence class of γ depends only on those of $\gamma_1,\dots,\gamma_r$.

Proof From (4-29) one gets

$$\sum_d \mu^d_{\mathcal{C}}(\gamma,\dots,\gamma) = \sum_k P^{r,k}_{\mathcal{C}}\Bigl(\sum_j \mu^j_{\mathcal{C}}(\gamma_k,\dots,\gamma_k)\Bigr). \tag{4-38}$$

Here, the P operations are defined using $\tilde{\gamma}_k = \gamma_k$. This shows that the Maurer–Cartan property is preserved. Similarly, suppose that for some $1 \leq k \leq r$, we have another Maurer–Cartan solution $\tilde{\gamma}_k$. Then, for the associated γ and $\tilde{\gamma} = \Pi^r_{\mathcal{C}}(\gamma_1,\dots,\tilde{\gamma}_k,\dots,\gamma_r)$,

$$\sum_{p,q} \mu^{p+q+1}_{\mathcal{C}}(\overbrace{\gamma,\dots,\gamma}^{p}, P^{r,k}_{\mathcal{C}}(x), \overbrace{\tilde{\gamma},\dots,\tilde{\gamma}}^{q}) = P^{r,k}_{\mathcal{C}}\Bigl(\sum_{p,q} \mu^{p+q+1}_{\mathcal{C}}(\overbrace{\gamma_k,\dots,\gamma_k}^{p}, x, \overbrace{\tilde{\gamma}_k,\dots,\tilde{\gamma}_k}^{q})\Bigr). \tag{4-39}$$

In particular, if we have an element h_k which provides an equivalence between γ_k and $\tilde{\gamma}_k$, then $h = P^{r,k}_{\mathcal{C}}(h_k)$ provides an equivalence between γ and $\tilde{\gamma}$, by (4-37). □

[4] The basic case is $\gamma_k = \tilde{\gamma}_k$, but for some applications, the freedom to choose a general $\tilde{\gamma}_k$ is important.

We want to mention a few elementary statements which, taken together, stand in a converse relation of sorts to Lemma 4.11.

Lemma 4.12 *Suppose that we have $\gamma_1, \dots, \gamma_{k-1}, \gamma_{k+1}, \dots, \gamma_r \in \mathcal{C}^1 \mathbin{\hat{\otimes}} N$. Then, for each $\gamma \in \mathcal{C}^1 \mathbin{\hat{\otimes}} N$, there is exactly one γ_k such that $\Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r) = \gamma$.*

Proof By (4-35) and (4-36), if $\xi \in \mathcal{C}^1 \hat{\otimes} N^m$, then $\Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_k + \xi, \dots, \gamma_r) = \Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r) + \xi \bmod N^{m+1}$. This allows one to solve for γ_k order by order, and to show uniqueness of the solution in the same way. □

Lemma 4.13 *Suppose that we have $\gamma_1, \dots, \gamma_r \in \mathcal{C}^1 \mathbin{\hat{\otimes}} N$. If all but γ_k are Maurer–Cartan elements, and $\gamma = \Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r)$ is Maurer–Cartan as well, then γ_k must also be Maurer–Cartan.*

Proof From (4-38) and the assumptions, one sees that $P^{r,k}_{\mathcal{C}}\bigl(\sum_j \mu^j_{\mathcal{C}}(\gamma_k, \dots, \gamma_k)\bigr) = 0$. On the other hand, by (4-36), $P^{r,k}_{\mathcal{C}}$ is clearly invertible. □

Lemma 4.14 *Given Maurer–Cartan elements $\gamma_1, \dots, \gamma_{k-1}, \gamma_{k+1}, \dots, \gamma_r$ and γ, there is a unique Maurer–Cartan element γ_k such that $\Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r) = \gamma$.*

Proof This is simply a combination of Lemmas 4.12 and 4.13. □

Lemma 4.15 *Suppose that we have Maurer–Cartan elements $\gamma_1, \dots, \gamma_r$ and $\tilde{\gamma}_k$ for some $1 \le k \le r$. If $\gamma = \Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r)$ and $\tilde{\gamma} = \Pi^r_{\mathcal{C}}(\gamma_1\ \dots, \gamma_{k-1}, \tilde{\gamma}_k, \gamma_{k+1}, \dots, \gamma_r)$ are equivalent, then so are γ_k and $\tilde{\gamma}_k$.*

Proof This is a consequence of (4-39) and the fact that $P^{r,k}_{\mathcal{C}}$ is an automorphism. □

Take the case $r = 1$ of (4-32). Then (4-29) says that $(\beta^1_{\mathcal{C}} = \mathrm{id}, \beta^2_{\mathcal{C}}, \dots)$ form an A_∞–homomorphism from $\mathcal{C}$ to itself (which is not surprising, since the underlying spaces MWW_d are the multiplihedra). The corresponding operation (4-32) is just the action of the A_∞–homomorphism on Maurer–Cartan elements. One can show that this A_∞–homomorphism is always homotopic to the identity, and hence $\Pi^1_{\mathcal{C}}(\gamma)$ is equivalent to γ. (The first piece of the statement about the A_∞–homomorphism is Example 4.8, but we won't explain the rest here; as for the action on Maurer–Cartan elements, we will give an indirect argument in Corollary 4.25.) Therefore, that case is essentially trivial. With that in mind, the first nontrivial instance of (4-32) is $r = 2$, which we will denote by

$$\gamma_1 \bullet \gamma_2 = \Pi^2_{\mathcal{C}}(\gamma_1, \gamma_2). \tag{4-40}$$

It will eventually turn out that the $r > 2$ cases can be reduced to an $(r-1)$–fold application of this product (Corollary 4.24), and hence are in a sense redundant.

4d Well-definedness

Proving that (4-32) is well-defined involves comparing different choices of the underlying A_∞–structures $\mu_{\mathcal{C}}$, as well as of the operations $\beta_{\mathcal{C}}$. Since the details are lengthy, and the outcome overall not surprising, we will provide only a sketch of the argument.

One can generalize the construction of the operations (4-26) by allowing the use of different versions of the A_∞–structure (in fact, a different version for each color of input, and another one for the output). Concretely, suppose that we have $(r+1)$ choices of fundamental chains for the Stasheff associahedra, with their associated A_∞–structures $\mu_{\mathcal{C},0}, \dots, \mu_{\mathcal{C},r}$. By choosing fundamental chains on the colored multiplihedra which satisfy an appropriately modified version of (4-24), we get generalized operations (4-26), which then lead to a map

$$\Pi^r_{\mathcal{C}} \colon \mathit{MC}(\mathcal{C}, \mu_{\mathcal{C},1}; N) \times \dots \times \mathit{MC}(\mathcal{C}, \mu_{\mathcal{C},r}; N) \to \mathit{MC}(\mathcal{C}, \mu_{\mathcal{C},0}; N). \tag{4-41}$$

For instance, let's look at $r = 1$. Then, what we get from the modified operations (4-26) is an A_∞–homomorphism between two choices of A_∞–structures on $\mathcal{C}$, whose linear part is the identity. That gives an alternative proof of the uniqueness result from Section 4b. (In spite of that, it made sense for us to include the original proof; the reason will become clear shortly.)

In (4-41), we want to understand the effect of simultaneously changing $\mu_{\mathcal{C},0}$, one of the other $\mu_{\mathcal{C},k+1}$, with $k \geq 0$, and correspondingly also (4-41). Namely, suppose that we have alternative versions $\tilde{\mu}_{\mathcal{C},0}$ and $\tilde{\mu}_{\mathcal{C},k+1}$. Alongside (4-41), we also have another operation which uses the alternative A_∞–structures, as well as different choices of functions (3-17) and fundamental chains on the MWW spaces. Let's denote that version by $\tilde{\Pi}^r_{\mathcal{C}}$. The construction from Section 4b yields A_∞–structures $\mu_{\mathcal{H},0}$ and $\mu_{\mathcal{H},k+1}$, where $\mathcal{H} = \mathcal{C} \otimes \mathcal{I}$. One can then construct a new operation $\Pi^{k,1,l}_{\mathcal{H}}$, where $k + 1 + l = r$, which fits into the following diagram, with vertical arrows induced by (2-5):

$$
\begin{tikzcd}
\begin{array}{r} \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},1}; N) \times \cdots \times \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},k+1}; N) \\ \times \cdots \times \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},r}; N) \end{array} \arrow[r, "\Pi^r_{\mathcal{C}}"] & \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},0}; N) \\
\begin{array}{r} \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},1}; N) \times \cdots \times \mathrm{MC}(\mathcal{H}, \mu_{\mathcal{H},k+1}; N) \\ \times \cdots \times \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},r}; N) \end{array} \arrow[u, "\cong"] \arrow[d, "\cong"'] \arrow[r, "\Pi^{k,1,l}_{\mathcal{H}}"] & \mathrm{MC}(\mathcal{H}, \mu_{\mathcal{H},0}; N) \arrow[u, "\cong"'] \arrow[d, "\cong"] \\
\begin{array}{r} \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},1}; N) \times \cdots \times \mathrm{MC}(\mathcal{C}, \tilde{\mu}_{\mathcal{C},k+1}; N) \\ \times \cdots \times \mathrm{MC}(\mathcal{C}, \mu_{\mathcal{C},r}; N) \end{array} \arrow[r, "\tilde{\Pi}^r_{\mathcal{C}}"] & \mathrm{MC}(\mathcal{C}, \tilde{\mu}_{\mathcal{C},0}; N)
\end{tikzcd}
\tag{4-42}
$$

Rather than giving the general construction of (4-42), we will only look at the $r = 1$ case. This is not terribly interesting in itself, but contains the main complications of the general situation, while allowing us to couch the discussion in more familiar terms. The setup for $r = 1$ is that we are given the following data:

- Four A_∞–structures on $\mathcal{C}$, namely $\mu_{\mathcal{C},k}$ and $\tilde{\mu}_{\mathcal{C},k}$ for $k = 0, 1$.
- Two A_∞–structures on $\mathcal{H}$, namely $\mu_{\mathcal{H},k}$ for $k = 0, 1$, which are constructed with the aim of interpolating between $\mu_{\mathcal{C},k}$ and $\tilde{\mu}_{\mathcal{C},k}$. Their definition, following (4-20), involves operations $\phi_{\mathcal{C},k}$ and $\psi_{\mathcal{C},k}$ as in (4-9) and (4-12).
- Finally, we have two versions of (4-26), which are A_∞–homomorphisms

$$
\begin{aligned}
\beta_{\mathcal{C}} &\colon (\mathcal{C}, \mu_{\mathcal{C},1}) \to (\mathcal{C}, \mu_{\mathcal{C},0}), \\
\tilde{\beta}_{\mathcal{C}} &\colon (\mathcal{C}, \tilde{\mu}_{\mathcal{C},1}) \to (\mathcal{C}, \tilde{\mu}_{\mathcal{C},0}).
\end{aligned}
$$

The aim is to define an A_∞–homomorphism $\beta_{\mathcal{H}}$, again having the identity as its linear term, which fits into a commutative diagram

$$
\begin{tikzcd}
(\mathcal{C}, \mu_{\mathcal{C},1}) \arrow[r, "\beta_{\mathcal{C}}"] & (\mathcal{C}, \tilde{\mu}_{\mathcal{C},0}) \\
(\mathcal{H}, \mu_{\mathcal{H},1}) \arrow[u] \arrow[d] \arrow[r, "\beta_{\mathcal{H}}"] & (\mathcal{H}, \mu_{\mathcal{H},0}) \arrow[u] \arrow[d] \\
(\mathcal{C}, \tilde{\mu}_{\mathcal{C},1}) \arrow[r, "\tilde{\beta}_{\mathcal{C}}"] & (\mathcal{C}, \tilde{\mu}_{\mathcal{C},0})
\end{tikzcd}
\tag{4-43}
$$

The corresponding special case of (4-42) is then defined through the action of $\beta_{\mathcal{H}}$ on Maurer–Cartan elements. The definition of $\beta_{\mathcal{H}}$ involves two kinds of operations:

$$\sigma_{\mathcal{C}}^{p,1,q}\colon \mathcal{C}^{\otimes p+1+q} \to \mathcal{C}[-p-q] \quad \text{for } p+q \geq 0, \tag{4-44}$$

with $\sigma_{\mathcal{C}}^{p,1,0} = \beta_{\mathcal{C}}^{p+1}$, $\sigma_{\mathcal{C}}^{0,1,q} = \widetilde{\beta}_{\mathcal{C}}^{q+1}$ and, in particular, $\sigma_{\mathcal{C}}^{0,1,0} = \mathrm{id}_{\mathcal{C}}$, and

$$\tau_{\mathcal{C}}^{p,q}\colon \mathcal{C}^{\otimes p+q} \to \mathcal{C}[-p-q-1] \quad \text{for } p,q > 0. \tag{4-45}$$

These enter into a formula parallel to equation (4-20): the nonzero terms of our A_∞–homomorphism are, for $p,q>0$,

$$\begin{aligned}
&\beta_{\mathcal{H}}^d(c_1 \otimes u, \dots, c_d \otimes u) = \beta_{\mathcal{C}}^d(c_1,\dots,c_d) \otimes u,\\
&\beta_{\mathcal{H}}^{p+q}(c_1 \otimes u, \dots, c_p \otimes u, c_{p+1} \otimes \widetilde{u}, \dots, c_{p+q} \otimes \widetilde{u})\\
&\qquad\qquad\qquad\qquad = (-1)^{\maltese_{p+q}} \tau_{\mathcal{C}}^{p,q}(c_1,\dots,c_p; c_{p+1},\dots,c_{p+q}) \otimes v,\\
&\beta_{\mathcal{H}}^d(c_1 \otimes \widetilde{u}, \dots, c_d \otimes \widetilde{u}) = \widetilde{\beta}_{\mathcal{C}}^d(c_1,\dots,c_d) \otimes \widetilde{u},\\
&\beta_{\mathcal{H}}^{p+q+1}(c_1 \otimes u, \dots, c_p \otimes u, c_{p+1} \otimes v, c_{p+2} \otimes \widetilde{u}, \dots, c_{p+q+1} \otimes \widetilde{u})\\
&\qquad = (-1)^{\maltese_{p+q+1} - \maltese_{p+1}} \sigma_{\mathcal{C}}^{p,1,q}(c_1,\dots,c_p; c_{p+1}; c_{p+2},\dots,c_{p+q+1}) \otimes v.
\end{aligned} \tag{4-46}$$

The fact that (4-46) satisfies the A_∞–homomorphism relations reduces to certain properties of (4-44) and (4-45). Those for (4-44) are

$$\begin{aligned}
&\sum_{p \geq i+j} (-1)^{\maltese_i} \sigma_{\mathcal{C}}^{p-j+1,1,q}\big(c_1,\dots,c_i,\mu_{\mathcal{C},1}^j(c_{i+1},\dots,c_{i+j}),\dots; c_{p+1};\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \dots, c_{p+q+1}\big)\\
&\quad + \sum_{i \leq p \leq i+j} (-1)^{\maltese_i} \sigma_{\mathcal{C}}^{i,1,p+q-i+j}\big(c_1,\dots,c_i;\\
&\qquad\qquad\qquad \phi_{\mathcal{C},1}^{p-i,1,i+j-p-1}(c_{i+1},\dots; c_{p+1}; \dots, c_{i+j});\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad c_{i+j+1},\dots,c_{p+q+1}\big)\\
&\qquad + \sum_{p<i} (-1)^{\maltese_i} \sigma_{\mathcal{C}}^{p,1,q-j+1}(c_1,\dots; c_{p+1};\\
&\qquad\qquad\qquad\qquad \dots, \widetilde{\mu}_{\mathcal{C},1}^j(c_{i+1},\dots,c_{i+j}),\dots,c_{p+q+1})\\
&= \sum_{\text{partitions}} \phi_{\mathcal{C},0}^{s,1,t}\big(\beta_{\mathcal{C}}^{d_1}(c_1,\dots,c_{d_1}),\dots,\beta_{\mathcal{C}}^{d_s}(c_{d_1+\dots+d_{s-1}+1},\dots,c_{d_1+\dots+d_s});\\
&\qquad\qquad \sigma_{\mathcal{C}}^{p-d_1-\dots-d_s,1,d_1+\dots+d_{s+1}-p-1}(c_{d_1+\dots+d_s+1},\dots; c_{p+1};\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad \dots, c_{d_1+\dots+d_{s+1}});\\
&\qquad\qquad\qquad \dots, \widetilde{\beta}_{\mathcal{C}}^{d_{s+1+t}}(c_{p+q+2-d_{s+t+1}},\dots,c_{p+q+1})\big).
\end{aligned} \tag{4-47}$$

On the right-hand side, the sum is over all (s,t) and partitions $d_1+\dots+d_{s+1+t} = p+1+q$ such that $d_1+\dots+d_s < p+1$ and $d_1+\dots+d_{s+1} \geq p+1$. In spite of the apparently larger number of terms which appear, this is formally parallel to the

A_∞–homomorphism equation, and in fact the nontrivial operations (4-44) are obtained from a choice of fundamental chains on MWW_{p+q+1}, as well as functions (3-17). The trick is that the boundary behavior of these data is partially determined by the choices underlying $\beta_{\mathcal{C}}$ and $\widetilde{\beta}_{\mathcal{C}}$, just as in our previous discussion of (4-8). Introducing the shorthand notation $\Sigma d_k^l := d_k + d_{k+1} + \cdots + d_l$ for $k < l$, the relations for (4-45) are

$$
\begin{aligned}
(4\text{-}48)\quad & -\sum_{p\geq i+j} (-1)^{\maltese_i} \tau_{\mathcal{C}}^{p-j+1,q}\big(c_1,\dots,c_i,\mu_{\mathcal{C},1}^j(c_{i+1},\dots,c_{i+j}),\dots; \\
&\qquad\qquad c_{p+1},\dots,c_{p+q}\big) \\
&+\sum_{i<p<i+j} \sigma_{\mathcal{C}}^{i,1,p+q-i-j}\big(c_1,\dots,c_i;\psi_{\mathcal{C},1}^{p-i,i+j-p}(c_{i+1},\dots;c_{p+1},\dots,c_{i+j}); \\
&\qquad\qquad c_{i+j+1},\dots,c_{p+q}\big) \\
&-\sum_{p\leq i}(-1)^{\maltese_i} \tau_{\mathcal{C}}^{p,q-j+1}(c_1,\dots,c_p;\dots,\widetilde{\mu}_{\mathcal{C},1}^j(c_{i+1},\dots,c_{i+j}),\dots,c_{p+q}) \\
&= \sum_{\text{partitions}} (-1)^{\maltese_{\Sigma d_1^s}} \phi_{\mathcal{C},0}^{s,1,t}\big(\beta_{\mathcal{C}}^{d_1}(c_1,\dots,c_{d_1}),\dots,\beta_{\mathcal{C}}^{d_s}(c_{\Sigma d_1^{s-1}+1},\dots,c_{\Sigma d_1^s}); \\
&\qquad\qquad \tau_{\mathcal{C}}^{p-\Sigma d_1^s,\Sigma d_1^{s+1}-p}(c_{\Sigma d_1^s+1},\dots,c_p;c_{p+1},\dots,c_{\Sigma d_1^{s+1}}); \\
&\qquad\qquad \dots,\widetilde{\beta}_{\mathcal{C}}^{d_{s+1+t}}(c_{p+q+2-d_{s+t+1}},\dots,c_{p+q})\big) \\
&\quad+\sum_{\text{partitions}} \psi_{\mathcal{C},0}^{s,t}\big(\beta_{\mathcal{C}}^{d_1}(c_1,\dots,c_{d_1}),\dots,\beta_{\mathcal{C}}^{d_s}(c_{\Sigma d_1^{s-1}+1},\dots,c_{\Sigma d_1^s}); \\
&\qquad\qquad \dots,\widetilde{\beta}_{\mathcal{C}}^{d_{s+t}}(c_{p+q+2-d_{s+t+1}},\dots,c_{p+q})\big).
\end{aligned}
$$

Combinatorially, the difference between the two terms on the right-hand side of (4-48) is where the dividing semicolon between the first p and last q inputs comes to lie: in the first case, we require that $\Sigma d_1^s := d_1 + \cdots + d_s < p < d_1 + \cdots + d_{s+1} =: \Sigma d_1^{s+1}$, so that semicolon is inside one of the innermost operations, which becomes a τ operation; in the second case, we require that $\Sigma d_1^s := d_1 + \cdots + d_s = p$, so that semicolon separates the two kinds of inputs for the ψ operation. Topologically one realizes (4-45) by choosing suitable fundamental chains on $[0,1]\times\mathrm{MWW}_{p+q}$, and analogues of (3-17) on that product space. The second sum in (4-48) contains terms which correspond to the boundary faces $\{0,1\}\times\mathrm{MWW}_{p+q}$, just as in (4-14):

$$
\begin{aligned}
(4\text{-}49)\quad & \sigma_{\mathcal{C}}^{p-1,1,q}(c_1,\dots,c_{p-1};\psi_{\mathcal{C},1}^{1,0}(c_p);c_{p+1},\dots,c_{p+q+1}) \\
&\qquad\qquad + \sigma_{\mathcal{C}}^{p,1,q-1}(c_1,\dots,c_p;\psi_{\mathcal{C},1}^{0,1}(c_p);c_{p+1},\dots,c_{p+q+1}) \\
&= -\sigma_{\mathcal{C}}^{p-1,1,q}(c_1,\dots,c_{p-1};c_p;c_{p+1},\dots,c_{p+q}) \\
&\qquad\qquad + \sigma_{\mathcal{C}}^{p,1,q-1}(c_1,\dots,c_p;c_{p+1};c_{p+2},\dots,c_{p+q}).
\end{aligned}
$$

Example 4.16 The first new operation $\tau_{\mathcal{C}}^{1,1}$ satisfies

$$\begin{aligned}(4\text{-}50)\quad \mu_{\mathcal{C}}^1(\tau_{\mathcal{C}}^{1,1}(c_1;c_2)) + \tau_{\mathcal{C}}^{1,1}(\mu_{\mathcal{C}}^1(c_1);c_2) + (-1)^{\|c_1\|}\tau_{\mathcal{C}}^{1,1}(c_1;\mu_{\mathcal{C}}^1(c_2)) \\ = \beta_{\mathcal{C}}^2(c_1,c_2) - \tilde{\beta}_{\mathcal{C}}^2(c_1,c_2) + \psi_{\mathcal{C},1}^{1,1}(c_1;c_2) - \psi_{\mathcal{C},0}^{1,1}(c_1;c_2),\end{aligned}$$

bearing in mind that all versions of the A_∞–structure on $\mathcal{C}$ share the same differential $-d = \mu_{\mathcal{C}}^1 = \phi_{\mathcal{C},0}^{0,1,0} = \phi_{\mathcal{C},1}^{0,1,0}$. This is exactly what's required for the first nontrivial A_∞–functor equation on $\mathcal{H}$: one has

$$\begin{aligned}(4\text{-}51)\quad \mu_{\mathcal{H},0}^1(\beta_{\mathcal{H}}^2(c_1\otimes u, c_2\otimes \tilde{u})) + \mu_{\mathcal{H},0}^2(\beta_{\mathcal{H}}^1(c_1\otimes u), \beta_{\mathcal{H}}^1(c_2\otimes \tilde{u})) \\ = (-1)^{\|c_1\|+\|c_2\|}\big(\mu_{\mathcal{C}}^1(\tau_{\mathcal{C}}^{1,1}(c_1;c_2)) + \psi_{\mathcal{C},0}^{1,1}(c_1;c_2)\big)\otimes v,\end{aligned}$$

while

$$\begin{aligned}(4\text{-}52)\quad \beta_{\mathcal{H}}^2(\mu_{\mathcal{H},1}^1(c_1\otimes u), c_2\otimes\tilde{u}) + (-1)^{\|c_1\|}\beta_{\mathcal{H}}^2(c_1\otimes u, \mu_{\mathcal{H},1}^1(c_2\otimes\tilde{u})) \\ +\beta_{\mathcal{H}}^1(\mu_{\mathcal{H},1}^2(c_1\otimes u, c_2\otimes\tilde{u})) \\ = (-1)^{\|c_1\|+\|c_2\|}\big(-\tau_{\mathcal{C}}^{1,1}(\mu_{\mathcal{C}}^1(c_1);c_2) + \beta_{\mathcal{C}}^2(c_1,c_2) - (-1)^{\|c_1\|}\tau_{\mathcal{C}}^{1,1}(c_1;\mu_{\mathcal{C}}^1(c_2)) \\ -\tilde{\beta}_{\mathcal{C}}^2(c_1,c_2) + \psi_{\mathcal{C},1}^{1,1}(c_1;c_2)\big)\otimes v.\end{aligned}$$

To conclude our discussion, let's return to the general context (arbitrary r), and note that then, repeated application of (4-42) allows one to change all the choices involved. We record the outcome:

Corollary 4.17 *Suppose that we have two different choices of fundamental chains on the associahedra and colored multiplihedra, as well as of functions* (3-17), *leading to two versions of the A_∞–structure and operations* (4-32). *These fit into a commutative diagram*

$$(4\text{-}53)\quad \begin{tikzcd} MC(\mathcal{C},\mu_{\mathcal{C}};N)^r \arrow[r, "\Pi_{\mathcal{C}}^r"] \arrow[d, "\cong"'] & MC(\mathcal{C},\mu_{\mathcal{C}};N) \arrow[d, "\cong"] \\ MC(\mathcal{C},\tilde{\mu}_{\mathcal{C}};N)^r \arrow[r, "\tilde{\Pi}_{\mathcal{C}}^r"] & MC(\mathcal{C},\tilde{\mu}_{\mathcal{C}};N) \end{tikzcd}$$

Here, we have related our A_∞–structures using functors as in (4-23), *and the vertical arrows are the induced maps on Maurer–Cartan elements.*

4e The p^{th} power operation

When defining (4-26), suppose now that we choose our functions (3-15) so that they satisfy (3-19). For the fundamental chains, we may also assume that they are chosen to be compatible with the identifications (3-18). In algebraic terms, the outcome is a cancellation property, which allows one to forget colors that do not carry any marked points:

$$\beta_{\mathcal{C}}^{d_1,\dots,d_{k-1},0,d_{k+1},\dots,d_r} = \beta_{\mathcal{C}}^{d_1,\dots,d_{k-1},d_{k+1},\dots,d_r}. \tag{4-54}$$

Assuming that such a choice has been adopted, we have:

Lemma 4.18 *Take a prime number p, and the coefficient ring $N = q\mathbb{F}_p[\![q]\!]/q^{p+1}$. Then, for $\gamma = qc + O(q^2)$, one has*

$$\Pi_{\mathcal{C}}^{p}(\gamma,\dots,\gamma) = \beta_{\mathcal{C}}^{1,\dots,1}(c;\dots;c)q^p. \tag{4-55}$$

Proof This is elementary, along the same lines as in Lemma 2.10. Applying (4-54) allows one to rewrite (4-32) as

$$\Pi_{\mathcal{C}}^{r}(\gamma,\dots,\gamma) = \sum_{1\le k\le r} \binom{r}{k} \sum_{d_1,\dots,d_k>0} \beta_{\mathcal{C}}^{d_1,\dots,d_k}(\overbrace{\gamma,\dots,\gamma}^{d_1};\dots;\overbrace{\gamma,\dots,\gamma}^{d_k}), \tag{4-56}$$

where the combinatorial factor reflects the possibilities of inserting 0 superscripts into each β operation. Suppose that our coefficient ring is $N = q\mathbb{F}_p[\![q]\!]$, and set $r = p$. Then (4-56) becomes

$$\Pi_{\mathcal{C}}^{p}(\gamma,\dots,\gamma) = \sum_{d_1,\dots,d_p>0} \beta_{\mathcal{C}}^{d_1,\dots,d_k}(\overbrace{\gamma,\dots,\gamma}^{d_1};\dots;\overbrace{\gamma,\dots,\gamma}^{d_p}). \tag{4-57}$$

Truncating mod q^{p+1} leaves

$$\beta_{\mathcal{C}}^{1,\dots,1}(\gamma;\dots;\gamma) = \beta_{\mathcal{C}}^{1,\dots,1}(c;\dots;c)q^p$$

as the only nonzero term. □

4f Deligne–Mumford spaces and commutativity

Let's consider the question of commutativity of the product (4-40). Concretely, this hypothetical commutativity would mean that there is an $h \in \mathcal{C}^0 \mathbin{\hat{\otimes}} N$ such that

$$\sum_{p,q} \mu_{\mathcal{C}}^{p+q+1}(\overbrace{\gamma_1 \bullet \gamma_2,\dots,\gamma_1 \bullet \gamma_2}^{p}, h, \overbrace{\gamma_2 \bullet \gamma_1,\dots,\gamma_2 \bullet \gamma_1}^{q}) = \gamma_1 \bullet \gamma_2 - \gamma_2 \bullet \gamma_1. \tag{4-58}$$

Let's suppose, to simplify the exposition, that the coefficient ring is $N = q\mathbb{Z}[\![q]\!]$. Moreover, we choose to define the operations (4-26) as in Section 4e, so that (4-54) holds. That entails some convenient (but not essential, of course) cancellations in our formulae. Given two Maurer–Cartan elements

$$\gamma_k = qc_k + O(q^2) \quad \text{for } k = 1, 2, \tag{4-59}$$

with leading terms c_k which are cocycles in $\mathcal{C}^1$, we have by definition,

$$\gamma_1 \bullet \gamma_2 - \gamma_2 \bullet \gamma_1 = q^2\big(\beta^{1,1}_{\mathcal{C}}(c_1; c_2) - \beta^{1,1}_{\mathcal{C}}(c_2; c_1)\big) + O(q^3). \tag{4-60}$$

In writing down this formula, we have exploited the fact that, due to our choices, $\beta^{2,0}_{\mathcal{C}}(c_k, c_k) = \beta^{0,2}_{\mathcal{C}}(c_k, c_k)$. It follows from Example 4.9 that

$$(c_1, c_2) \in \mathcal{C} \mapsto (-1)^{\|c_1\|}\beta^{1,1}_{\mathcal{C}}(c_1; c_2) - (-1)^{\|c_1\|\,|c_2|}\beta^{1,1}_{\mathcal{C}}(c_2; c_1) \tag{4-61}$$

is a chain map of degree -1. On cohomology, it defines the Lie bracket

$$[\cdot,\cdot]\colon H^*(\mathcal{C}) \otimes H^*(\mathcal{C}) \to H^{*-1}(\mathcal{C}),$$

which is part of the Gerstenhaber algebra structure. Geometrically, $\beta^{1,1}$ arises from a one-dimensional chain in FM_2 whose boundary points are exchanged by the $\mathbb{Z}/2$–action. The sum of this chain and its image under the nontrivial element of $\mathbb{Z}/2$ is a cycle, which generates $H_1(\mathrm{FM}_2) \cong H_1(S^1) = \mathbb{Z}$. If we similarly write $h = qb + O(q^2)$, then (4-58) taken modulo q^3 says that b is a cocycle, and that

$$\mu^2_{\mathcal{C}}(c_1+c_2, b) + \mu^2_{\mathcal{C}}(b, c_1+c_2) + (\text{coboundaries}) = \beta^{1,1}_{\mathcal{C}}(c_1; c_2) - \beta^{1,1}_{\mathcal{C}}(c_2; c_1). \tag{4-62}$$

By (4-31), the left-hand side of (4-62) is nullhomologous. Hence, for (4-62) to be satisfied, the Lie bracket of $[c_1]$ and $[c_2]$ must be zero, which means that commutativity does not hold in this level of generality.

We now switch from Fulton–MacPherson to Deligne–Mumford spaces. One could define the structure of an algebra over $C_{-*}(\mathrm{DM}_d)$ on a chain complex $\mathcal{C}$ in the same way as before. However, that notion is not well-behaved. For instance, the action of $\mathrm{DM}_2 = \text{point}$ would yield a strictly commutative product on $\mathcal{C}$. The underlying problem is that the Sym_d–action on DM_d is not free — from an algebraic viewpoint, $C_{-*}(\mathrm{DM}_d)$ is not a projective $\mathbb{Z}[\mathrm{Sym}_d]$–module. There is a simple workaround, by "freeing up" the action. Namely, let (E_d) be an E_∞–operad, which means that the spaces E_d are contractible and freely acted on by Sym_d. Let's adopt a concrete choice, namely, the analogue of Fulton–MacPherson space for point configurations in $\mathbb{R}^\infty$. Then

$$\mathcal{DM}_d = \mathrm{DM}_d \times E_d \tag{4-63}$$

is again an operad, which is homotopy equivalent to DM_d but carries a free action of Sym_d. The maps (3-8) admit lifts

$$\mathrm{FM}_d \to \mathcal{DM}_d \tag{4-64}$$

which are compatible with the operad structure, including the action of Sym_d. As an existence statement, this is a consequence of the properties of E_d; but for our specific choice, such lifts can be defined explicitly by taking (3-8) together with the natural inclusion $\mathrm{FM}_d \to E_d$.

Assume from now on that $\mathcal{C}$ carries the structure of an operad over $C_{-*}(\mathcal{DM}_d)$, and hence inherits one over the Fulton–MacPherson operad by (4-64). Take the one-cycle in FM_2 underlying (4-60) and map it to (the contractible space) $\mathcal{DM}_2$. Choosing a bounding cochain (which is itself unique up to coboundaries) yields a nullhomotopy

$$\begin{gathered}\kappa^{1,1}_{\mathcal{C}}\colon \mathcal{C}^{\otimes 2} \to \mathcal{C}[-2], \\ \mu^1_{\mathcal{C}}(\kappa^{1,1}_{\mathcal{C}}(c_1;c_2)) + \kappa^{1,1}_{\mathcal{C}}(\mu^1_{\mathcal{C}}(c_1);c_2) + (-1)^{\|c_1\|}\kappa^{1,1}_{\mathcal{C}}(c_1;\mu^1_{\mathcal{C}}(c_2)) \\ = \beta^{1,1}_{\mathcal{C}}(c_1;c_2) - (-1)^{\|c_1\|\,\|c_2\|}\beta^{1,1}_{\mathcal{C}}(c_2;c_1).\end{gathered} \tag{4-65}$$

As a consequence, if we set

$$h = q^2\kappa^{1,1}_{\mathcal{C}}(c_1;c_2) \in \mathcal{C}^0 \mathbin{\hat{\otimes}} q^2\mathbb{Z}[\![q]\!], \tag{4-66}$$

then (4-58) is satisfied modulo q^3 (on the left-hand side, only the $p=q=0$ term matters at this point). Hence, if we reduce coefficients to $q\mathbb{Z}[\![q]\!]/q^3$, then (4-40) is commutative. Nothing we have said so far is in any way surprising: the vanishing of the Lie bracket in the case where the operations come from Deligne–Mumford space is a well-known fact — if one uses the framed little disc operad as an intermediate object, it follows from vanishing of the BV operator.

Let's push our investigation a little further. As special cases of (4-29), we have

$$\begin{aligned}&\mu^1_{\mathcal{C}}(\beta^{2,1}_{\mathcal{C}}(c_1,c_2;c_3)) - \beta^{2,1}_{\mathcal{C}}(\mu^1_{\mathcal{C}}(c_1),c_2;c_3) \\ &\qquad - (-1)^{\|c_1\|}\beta^{2,1}_{\mathcal{C}}(c_1,\mu^1_{\mathcal{C}}(c_2);c_3) - (-1)^{\|c_1\|+\|c_2\|}\beta^{2,1}_{\mathcal{C}}(c_1,c_2;\mu^1_{\mathcal{C}}(c_3)) \\ &= \beta^{1,1}_{\mathcal{C}}(\mu^2_{\mathcal{C}}(c_1,c_2);c_3) - (-1)^{\|c_2\|\,\|c_3\|}\mu^2_{\mathcal{C}}(\beta^{1,1}_{\mathcal{C}}(c_1;c_3),c_2) \\ &\qquad - \mu^2_{\mathcal{C}}(c_1,\beta^{1,1}_{\mathcal{C}}(c_2;c_3)) - \mu^2_{\mathcal{C}}(\beta^2_{\mathcal{C}}(c_1,c_2),c_3) \\ &\qquad\quad - (-1)^{\|c_3\|(\|c_1\|+\|c_2\|)}\mu^2_{\mathcal{C}}(c_3,\beta^2_{\mathcal{C}}(c_1,c_2)) - \mu^3_{\mathcal{C}}(c_1,c_2,c_3) \\ &\qquad\qquad - (-1)^{\|c_2\|\,\|c_3\|}\mu^3_{\mathcal{C}}(c_1,c_3,c_2) - (-1)^{(\|c_1\|+\|c_2\|)\|c_3\|}\mu^3_{\mathcal{C}}(c_3,c_1,c_2),\end{aligned} \tag{4-67}$$

and respectively,

$$\begin{aligned}
(4\text{-}68)\quad & \mu^1_{\mathcal{C}}(\beta^{1,2}_{\mathcal{C}}(c_3;c_1,c_2)) - \beta^{1,2}_{\mathcal{C}}(\mu^1_{\mathcal{C}}(c_3);c_1,c_2) \\
& \quad - (-1)^{\|c_3\|}\beta^{1,2}_{\mathcal{C}}(c_3;\mu^1_{\mathcal{C}}(c_1),c_2) - (-1)^{\|c_3\|+\|c_1\|}\beta^{1,2}_{\mathcal{C}}(c_3;c_1,\mu^1_{\mathcal{C}}(c_2)) \\
& = (-1)^{\|c_3\|}\beta^{1,1}_{\mathcal{C}}(c_3;\mu^2_{\mathcal{C}}(c_1,c_2)) - \mu^2_{\mathcal{C}}(\beta^{1,1}_{\mathcal{C}}(c_3;c_1),c_2) \\
& \quad - (-1)^{\|c_1\|\|c_3\|}\mu^2_{\mathcal{C}}(c_1,\beta^{1,1}_{\mathcal{C}}(c_3;c_2)) - \mu^2_{\mathcal{C}}(c_3,\beta^2_{\mathcal{C}}(c_1,c_2)) \\
& \qquad - (-1)^{\|c_3\|(\|c_1\|+\|c_2\|)}\mu^2_{\mathcal{C}}(\beta^2_{\mathcal{C}}(c_1,c_2),c_3) - \mu^3_{\mathcal{C}}(c_3,c_1,c_2) \\
& \qquad\quad - (-1)^{\|c_3\|\|c_1\|}\mu^3_{\mathcal{C}}(c_1,c_3,c_2) - (-1)^{\|c_3\|(\|c_1\|+\|c_2\|)}\mu^3_{\mathcal{C}}(c_1,c_2,c_3).
\end{aligned}$$

Therefore, if we consider the map $K^{2,1}_{\mathcal{C}}\colon \mathcal{C}^{\otimes 3} \to \mathcal{C}[-2]$ given by

$$\begin{aligned}
(4\text{-}69)\quad & K^{2,1}_{\mathcal{C}}(c_1,c_2;c_3) \\
& = \beta^{2,1}_{\mathcal{C}}(c_1,c_2;c_3) - (-1)^{\|c_3\|(\|c_1\|+\|c_2\|)}\beta^{1,2}_{\mathcal{C}}(c_3;c_1,c_2) - \kappa^{1,1}_{\mathcal{C}}(\mu^2_{\mathcal{C}}(c_1,c_2);c_3) \\
& \qquad - (-1)^{\|c_2\|\|c_3\|}\mu^2_{\mathcal{C}}(\kappa^{1,1}_{\mathcal{C}}(c_1;c_3),c_2) - (-1)^{\|c_1\|}\mu^2_{\mathcal{C}}(c_1,\kappa^{1,1}_{\mathcal{C}}(c_2;c_3)),
\end{aligned}$$

then that satisfies

$$\begin{aligned}
(4\text{-}70)\quad & \mu^1_{\mathcal{C}}(K^{2,1}_{\mathcal{C}}(c_1,c_2;c_3)) - K^{2,1}_{\mathcal{C}}(\mu^1_{\mathcal{C}}(c_1),c_2;c_3) \\
& \quad - (-1)^{\|c_1\|}K^{2,1}_{\mathcal{C}}(c_1,\mu^1_{\mathcal{C}}(c_2);c_3) - (-1)^{\|c_1\|+\|c_2\|}K^{2,1}_{\mathcal{C}}(c_1,c_2;\mu^1_{\mathcal{C}}(c_3)) = 0,
\end{aligned}$$

which means that $(-1)^{|c_2|}K^{2,1}_{\mathcal{C}}(c_1,c_2;c_3)$ is a chain map of degree -2. The same observation applies to

$$\begin{aligned}
(4\text{-}71)\quad & K^{1,2}_{\mathcal{C}}(c_1;c_2,c_3) \\
& = \beta^{1,2}_{\mathcal{C}}(c_1;c_2,c_3) - (-1)^{\|c_1\|(\|c_2\|+\|c_3\|)}\beta^{2,1}_{\mathcal{C}}(c_2,c_3;c_1) \\
& \qquad - (-1)^{\|c_1\|}\kappa^{1,1}_{\mathcal{C}}(c_1;\mu^2_{\mathcal{C}}(c_2,c_3)) - \mu^2_{\mathcal{C}}(\kappa^{1,1}_{\mathcal{C}}(c_1;c_2),c_3) \\
& \qquad\qquad - (-1)^{\|c_1\|\|c_2\|}\mu^2_{\mathcal{C}}(c_2,\kappa^{1,1}_{\mathcal{C}}(c_1;c_3)).
\end{aligned}$$

We now return to the original situation (4-59). Suppose that the cocycles $K^{2,1}_{\mathcal{C}}(c_1,c_1;c_2)$ and $K^{1,2}_{\mathcal{C}}(c_1;c_2,c_2)$ are trivial in $H^*(\mathcal{C})$, and that we have chosen bounding cochains for them,

$$(4\text{-}72)\qquad K^{2,1}_{\mathcal{C}}(c_1,c_1;c_2) = \mu^1_{\mathcal{C}}(b^{2,1}) \quad\text{and}\quad K^{1,2}_{\mathcal{C}}(c_1;c_2,c_2) = \mu^1_{\mathcal{C}}(b^{1,2}).$$

Then (4-58) is satisfied modulo q^4 by the refinement of (4-66) given by

$$(4\text{-}73)\qquad h = \kappa^{1,1}_{\mathcal{C}}(\gamma_1;\gamma_2) + q^3(b^{2,1}+b^{1,2}) \in \mathcal{C}^0 \mathbin{\hat{\otimes}} q\mathbb{Z}[\![q]\!]/q^4.$$

It remains to look at the geometry underlying (4-69) and (4-71). Both cases are parallel, so let's focus on $K^{2,1}_{\mathcal{C}}$. For $\beta^{2,1}_{\mathcal{C}}(c_1, c_2; c_3)$, take the relevant map and project it to actual Deligne–Mumford space for simplicity, which means considering the composition

$$\mathrm{MWW}_{2,1} \to \mathrm{FM}_3 \to \mathcal{DM}_3 \to \mathrm{DM}_3 \cong S^2. \tag{4-74}$$

Looking at Figure 5, we see that three boundary sides of the octagon $\mathrm{MWW}_{2,1}$, corresponding to the μ^3 terms in (4-69), are mapped to paths in DM_3 which are images of the canonical map $S_3 \to \mathrm{DM}_3$ and two of its permuted versions (those which preserve the ordering between the first and second point in the configuration). The remaining sides are collapsed to "special points", meaning the images of the maps $\mathrm{DM}_2 \times \mathrm{DM}_2 \to \mathrm{DM}_3$. Altogether, we get a relative cycle, whose homology class in $H_2(\mathrm{DM}_3, (\mathrm{DM}_3)_{\mathbb{R}}) = H_2(S^2, S^1)$ is independent of all choices involved in the construction. From the assumption (3-16), one sees easily that this cycle corresponds to one of the two discs in S^2 bounding S^1. Correspondingly, for $\beta^{1,2}_{\mathcal{C}}(c_3; c_1, c_2)$, we get a relative cycle corresponding to the other disc. The outcome of this discussion is that $K^{2,1}_{\mathcal{C}}$ is constructed from a cycle which represents a generator of $H_2(\mathrm{DM}_3) = H_2(S^2)$; roughly speaking, the difference between the two discs bounding the same S^1. Since the action of Sym_3 on $H_2(\mathrm{DM}_3)$ is trivial, the induced map

$$([c_1], [c_2], [c_3]) \mapsto [(-1)^{|c_2|} K^{2,1}_{\mathcal{C}}(c_1, c_2; c_3)] \colon H^*(\mathcal{C})^{\otimes 3} \to H^{*-2}(\mathcal{C}) \tag{4-75}$$

is graded symmetric. In particular, $[K^{2,1}_{\mathcal{C}}(c_1, c_1; c_2)] \in H^1(\mathcal{C})$ must be 2–torsion. If we can rule out such torsion, then the class must necessarily vanish, and similarly for $[K^{1,2}_{\mathcal{C}}(c_1; c_2, c_2)]$. We can carry over the argument to other coefficients:

Proposition 4.19 *The product $\bullet$ on $\mathrm{MC}(\mathcal{C}; N)$ is commutative if $N^3 = 0$. It is also commutative if $N^4 = 0$ and, additionally, $H^*(\mathcal{C})$ is a free abelian group.*

Proof The first part is as in (4-66). For the second part, the obstruction is now $[K^{2,1}_{\mathcal{C}}(c_1, c_1; c_2)] \in H^{*-2}(\mathcal{C}; N^3)$, and similarly for $K^{1,2}_{\mathcal{C}}$. From our assumption, it follows that $H^*(\mathcal{C}; G) = H^*(\mathcal{C}) \otimes G$ for any abelian group G. Hence, the symmetry argument that ensures vanishing of (4-75) carries over to arbitrary coefficients. □

4g Strip-shrinking spaces and associativity

Let's assume that $\mathcal{C}$ is homologically unital; this assumption is used in the context of geometric stabilization arguments, which add extra marked points. Our aim is to show:

Proposition 4.20 *For any $r \geq 2$ and any $1 \leq m \leq r-1$, $\Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r)$ is equivalent to $\Pi^{r-1}_{\mathcal{C}}(\gamma_1, \dots, \gamma_m \bullet \gamma_{m+1}, \dots, \gamma_r) = \Pi^{r-1}_{\mathcal{C}}(\gamma, \dots, \Pi^2_{\mathcal{C}}(\gamma_m, \gamma_{m+1}), \dots, \gamma_r)$.*

Before getting into the proof, let's draw some immediate consequences.

Corollary 4.21 *The product $\bullet$ is associative.*

This is because, by the $r = 3$ case of Proposition 4.20, $\Pi^3_{\mathcal{C}}(\gamma_1, \gamma_2, \gamma_3)$ is equivalent to both $\Pi^2_{\mathcal{C}}(\gamma_1 \bullet \gamma_2, \gamma_3) = (\gamma_1 \bullet \gamma_2) \bullet \gamma_3$ and $\Pi^2_{\mathcal{C}}(\gamma_1, \gamma_2 \bullet \gamma_3) = \gamma_1 \bullet (\gamma_2 \bullet \gamma_3)$.

Corollary 4.22 *The neutral element is $\gamma = 0$, meaning that $\gamma \bullet 0$ and $0 \bullet \gamma$ are both equivalent to γ.*

The definition says that

$$\gamma \bullet 0 = \sum_d \beta^{d,0}_{\mathcal{C}}(\gamma, \dots, \gamma), \tag{4-76}$$

so the statement is not immediately obvious. However, it is obvious that $0 \bullet 0 = 0$. By Lemmas 4.14 and 4.15, $\gamma \mapsto \gamma \bullet 0$ is a bijective map from $\mathit{MC}(\mathcal{C}; N)$ to itself. By associativity, that bijective map is idempotent, and therefore the identity. (There is a more direct geometric argument which shows that (4-76) is equivalent to γ, along the lines of Example 4.8, but we have chosen to avoid it.)

Corollary 4.23 $(\mathit{MC}(\mathcal{C}; N), \bullet)$ *is a group.*

This is a combination of the previous two corollaries and Lemma 4.14.

Corollary 4.24 *For any $r \geq 3$, $\Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r)$ is equivalent to $\gamma_1 \bullet \dots \bullet \gamma_r$.*

This follows by induction: if $\Pi^{r-1}_{\mathcal{C}}(\gamma_1, \dots, \gamma_{r-1})$ is equivalent to $\gamma_1 \bullet \dots \bullet \gamma_{r-1}$, then $\Pi^r_{\mathcal{C}}(\gamma_1, \dots, \gamma_r)$ is equivalent to $\Pi^{r-1}_{\mathcal{C}}(\gamma_1 \bullet \gamma_2, \dots, \gamma_r)$, hence to $(\gamma_1 \bullet \gamma_2) \bullet \dots \bullet \gamma_r$.

Corollary 4.25 $\Pi^1_{\mathcal{C}}(\gamma)$ *is always equivalent to γ.*

Proposition 4.20, with $r = 2$, says that $\Pi^2_{\mathcal{C}}(\gamma_1, \gamma_2) = \gamma_1 \bullet \gamma_2$ is equivalent to $\Pi^1_{\mathcal{C}}(\gamma_1 \bullet \gamma_2)$, which implies the desired statement by specializing to $\gamma_1 = 0$ (again, this is a workaround which avoids a direct geometric argument).

Proof of Proposition 4.20 This uses the strip-shrinking moduli spaces from Section 3f, with their maps (3-31). The codimension-one boundary faces are images of maps (3-22) defined on the following spaces. First, in parallel with the first term of (4-24), we have

$$\mathrm{SS}_{d_1,\dots,d_k-j+1,\dots,d_r} \times S_j, \tag{4-77}$$

where j and k can be arbitrary (in particular, the latter can be m or $m+1$, something that's not entirely reflected in our notation). The analogue of the second term in (4-24) is less obvious:

$$\begin{aligned} S_j \times \prod_{i=1}^{k-1} \overbrace{\mathrm{MWW}_{d_{i,1},\dots,d_{i,r}} \times \mathrm{SS}_{d_{k,1},\dots,(d_{k,m},d_{k,m+1}),\dots,d_{k,r}}}^{r \text{ colors}} \\ \times \prod_{i=k+1}^{j} \Big(\underbrace{\mathrm{MWW}_{d_{i,1},\dots,d_{i,m-1},a_i,d_{i,m+2},\dots,d_{i,r}}}_{r-1 \text{ colors}} \times \prod_{l=1}^{a_i} \underbrace{\mathrm{MWW}_{d_{i,l,1},d_{i,l,2}}}_{2 \text{ colors}} \Big). \end{aligned} \tag{4-78}$$

The last kind of MWW factor corresponds to the small-mid vertices in the terminology of (3-22). Such boundary faces are parametrized by "double partitions". One first chooses $j \geq 2$ and partitions

$$d_1 = d_{1,1} + \cdots + d_{j,1}, \quad \dots, \quad d_r = d_{1,r} + \cdots + d_{j,r}.$$

In addition, there is a distinguished $k \in \{1, \dots, j\}$ for which $(d_{k,1}, \dots, d_{k,r})$ can be $(0, \dots, 0)$. Finally, for each $i > k$, one chooses a further a_i and corresponding partitions $d_{i,m} = d_{i,1,1} + \cdots + d_{i,a_i,1}$ and $d_{i,m+1} = d_{i,1,2} + \cdots + d_{i,a_i,2}$.

We fix fundamental chains on the SS spaces, compatible with the boundary structure in the usual sense. We then insert those chains into our operadic structure through (3-31), with the additional convention that at the stabilizing marked points (3-30), we will always apply a fixed homology unit $e_{\mathcal{C}}$. Denote the resulting operations by

$$\alpha_{\mathcal{C}}^{d_1,\dots,(d_m,d_{m+1}),\dots,d_r} : \mathcal{C}^{\otimes d} \to \mathcal{C}[-d], \quad \text{where } d = d_1 + \cdots + d_r \geq 0. \tag{4-79}$$

The simplest example is $\mathrm{SS}_{0,\dots,0} = \text{point}$, which is mapped to $\mathrm{FM}_2 = S^1$ by taking the configuration (3-30). This coincides with the map $S_2 \to \mathrm{FM}_2$ which is part of our A_∞–structure, and therefore

$$\alpha_{\mathcal{C}}^{0,\dots,(0,0),\dots,0} = \mu_{\mathcal{C}}^2(e_{\mathcal{C}}, e_{\mathcal{C}}) = e_{\mathcal{C}} + (\text{coboundary}). \tag{4-80}$$

Generally, the operations (4-79) satisfy the equation obtained from setting the sum of boundary contributions (4-77), (4-78) equal to zero:

$$\begin{aligned}(4\text{-}81)\quad \sum_{ijk} \pm\alpha_{\mathcal{C}}^{d_1,\dots,d_k-j+1,\dots,d_r}\big(&c_{1,1},\dots,c_{1,d_1};\dots;\\ &c_{k,1},\dots,\mu_{\mathcal{C}}^{j}(c_{k,i+1},\dots,c_{k,i+j}),\dots,c_{k,d_k};\dots;\\ &c_{r,1},\dots,c_{r,d_r}\big)\end{aligned}$$

$$\begin{aligned}= \sum_{\substack{\text{double}\\ \text{partitions}}} \pm\mu_{\mathcal{C}}^{j}\Big(&\beta_{\mathcal{C}}^{d_{1,1},\dots,d_{1,r}}(c_{1,1},\dots,c_{1,d_{1,1}};\dots;c_{r,1},\dots,c_{r,d_{1,r}}),\dots,\\ &\alpha_{\mathcal{C}}^{d_{k,1},\dots,d_{k,r}}(c_{1,d_{1,1}+\dots+d_{k-1,1}+1},\dots,c_{1,d_{1,1}+\dots+d_{k,1}};\dots),\dots,\\ &\beta_{\mathcal{C}}^{d_{j,1},\dots,d_{j,m-1},a_j,d_{j,m+2},\dots,d_{j,r}}\big(c_{1,d_1-d_{j,1}+1},\dots,c_{1,d_1};\\ &\quad\beta_{\mathcal{C}}^{d_{j,1,1},d_{j,1,2}}(c_{m,d_m-d_{j,m}+1},\dots;c_{m+1,d_{m+1}-d_{j,m+1}+1},\dots),\\ &\quad\dots,\beta_{\mathcal{C}}^{d_{j,a_j,1},d_{j,a_j,2}}(\dots,c_{m,d_m};\dots,c_{m+1,d_{m+1}});\\ &\quad c_{m+2,d_{m+2}-d_{m+2,j}+1},\dots,c_{m+2,d_{m+2}};\dots\big)\Big),\end{aligned}$$

with double partitions as in (4-78), the only difference being that we have an additional $\mu_{\mathcal{C}}^1(\beta_{\mathcal{C}}^{d_1,\dots,d_r})$ term. Given Maurer–Cartan elements $\gamma_1,\dots,\gamma_r$, set

$$(4\text{-}82)\quad g = \sum_{d_1,\dots,d_r\geq 0} \alpha_{\mathcal{C}}^{d_1,\dots,(d_m,d_{m+1}),\dots,d_r}(\overbrace{\gamma_1,\dots,\gamma_1}^{d_1};\dots;\overbrace{\gamma_r,\dots,\gamma_r}^{d_r}) \in \mathcal{C}^0 \mathbin{\hat{\otimes}} (\mathbb{Z}1\oplus N).$$

As a direct consequence of (4-81), this satisfies

$$\begin{aligned}(4\text{-}83)\quad \sum_{p,q}\mu_{\mathcal{C}}^{p+q+1}\Big(&\overbrace{\Pi_{\mathcal{C}}^{r}(\gamma_1,\dots,\gamma_r),\dots,\Pi_{\mathcal{C}}^{r}(\gamma_1,\dots,\gamma_r)}^{p},g,\\ &\underbrace{\Pi_{\mathcal{C}}^{r-1}(\gamma_1,\dots,\gamma_m\bullet\gamma_{m+1},\dots,\gamma_r),\dots,\Pi_{\mathcal{C}}^{r-1}(\gamma_1,\dots,\gamma_m\bullet\gamma_{m+1},\dots,\gamma_r)}_{q}\Big)\\ &= 0.\end{aligned}$$

In view of (4-80) and Lemma 2.6, this is exactly what we need to prove the equivalence of the two Maurer–Cartan elements in question. □

5 Cohomology operations

Following Steenrod and (in a more abstract context) May, reduced power operations arise from homotopy symmetries. This general principle can be applied to configuration spaces, as in Cohen's classical work, and also to Deligne–Mumford spaces. After a brief review of the underlying homological algebra, we discuss those two instances, and their relationship.

5a Equivariant (co)homology

Let $\mathcal{C}$ be a complex of vector spaces over a field $\mathbb{F}$. Given an action of a group G on $\mathcal{C}$, one can consider the group cochain complex $C_G^*(\mathcal{C})$ and its cohomology $H_G^*(\mathcal{C})$. (This is a mild generalization of the classical concept of equivariant cohomology, where the coefficients lie in a G–module; see eg [10] for a general account, and [46, page 115] or [73, page 179] for the traditional choice of group cochain complex.) If $C^*(X)$ is the cochain complex of a space X carrying a G–action, and V is a representation of G over $\mathbb{F}$, then setting $\mathcal{C} = C^*(X) \otimes V$ recovers

$$H_G^*(C^*(X) \otimes V) = H_G^*(X; V) = H^*(X \times_G EG; V), \tag{5-1}$$

the equivariant cohomology in the classical sense (with coefficients in the local system over the Borel construction $X \times_G EG$ determined by V). A variant of the construction yields the group chain complex $C_*^G(\mathcal{C})$ and group homology $H_*^G(\mathcal{C})$. Recall that in our convention, all chain complexes are cohomologically graded. If we start with the chain complex $C_*(X)$ of a space, and a representation V, as before, then setting $\mathcal{C} = C_{-*}(X) \otimes V$ gives

$$H_*^G(C_{-*}(X) \otimes V) = H_{-*}^G(X; V) = H_{-*}(X \times_G EG; V). \tag{5-2}$$

Group homology and cohomology carry exterior cup and cap products

$$H_G^*(\mathcal{C}_1) \otimes H_G^*(\mathcal{C}_2) \to H_G^*(\mathcal{C}_1 \otimes \mathcal{C}_2), \tag{5-3}$$

$$H_*^G(\mathcal{C}_1) \otimes H_G^*(\mathcal{C}_2) \to H_*^G(\mathcal{C}_1 \otimes \mathcal{C}_2); \tag{5-4}$$

see eg [10, Section V.3]. The cases relevant for our purpose are where G is a permutation group Sym_p of prime order, or its cyclic subgroup $\mathbb{Z}/p$, and $\mathbb{F} = \mathbb{F}_p$. For the cyclic group, there are particularly simple complexes computing equivariant (co)homology. The cohomology version is

$$\begin{cases} C_{\mathbb{Z}/p}^*(\mathcal{C}) = \mathcal{C}[\![t]\!] \oplus \theta\mathcal{C}[\![t]\!], \quad \text{with } |t| = 2,\ |\theta| = 1, \\ d_{\mathbb{Z}/p}(t^k c) = t^k\, dc + \theta t^k (Tc - c), \\ d_{\mathbb{Z}/p}(\theta t^k c) = -\theta t^k\, dc + t^{k+1}(c + Tc + \cdots + T^{p-1}c), \end{cases} \tag{5-5}$$

where $T \colon \mathcal{C} \to \mathcal{C}$ is the generator of the $\mathbb{Z}/p$–action. In the case of trivial coefficients $\mathcal{C} = \mathbb{F}_p$, the differential vanishes. The ring structure (5-3), for $\mathcal{C}_1 = \mathbb{F}_p$ and general $\mathcal{C}_2 = \mathcal{C}$, satisfies $[t] \cdot [t^k c] = [t^{k+1} c]$ and $[t] \cdot [t^k \theta c] = [t^{k+1} \theta c]$. However, for $p = 2$ and $\mathcal{C}_1 = \mathcal{C}_2 = \mathbb{F}_p$, one has $[\theta] \cdot [\theta] = [t]$, while for $p > 2$ that expression would be zero. Indeed, in the case $p = 2$ it is more convenient to write $\theta = t^{1/2}$; for a more precise discussion of the choices of generators used here in relation to topology, we refer to Section 10a. The group homology version is

$$\text{(5-6)}\qquad \begin{cases} C_*^{\mathbb{Z}/p}(\mathcal{C}) = \mathcal{C}[s] \oplus \sigma\mathcal{C}[s], \quad \text{with } |s| = -2,\ |\sigma| = -1, \\ d^{\mathbb{Z}/p}(s^k c) = s^k\, dc - \sigma s^{k-1}(c + Tc + \cdots + T^{p-1}c), \\ d^{\mathbb{Z}/p}(\sigma s^k c) = -\sigma s^k\, dc - s^k(Tc - c), \end{cases}$$

and under (5-4), t acts on equivariant homology by canceling one power of s in (5-6) (by convention, s^{-1} is set to zero). Because the index of $\mathbb{Z}/p \subset \mathrm{Sym}_p$ is coprime to p, for every complex $\mathcal{C}$ with Sym_p–action,

$$\text{(5-7)}\qquad H^*_{\mathrm{Sym}_p}(\mathcal{C}) \to H^*_{\mathbb{Z}/p}(\mathcal{C}) \quad \text{is injective,}$$

$$\text{(5-8)}\qquad H_*^{\mathbb{Z}/p}(\mathcal{C}) \to H_*^{\mathrm{Sym}_p}(\mathcal{C}) \quad \text{is surjective.}$$

Let $\mathbb{F}_p(l)$ with $l \in \mathbb{Z}$ be the one-dimensional representations which are

- trivial if l is even, and
- associated to sign: $\mathrm{Sym}_p \to \{\pm 1\} \subset \mathbb{F}_p^\times$ if l is odd.

Note that the restriction of the sign homomorphism to $\mathbb{Z}/p$ is trivial. The relevant special case of (5-7) can be made explicit:

$$\text{(5-9)}\qquad H^*_{\mathrm{Sym}_p}(\mathbb{F}_p) = \mathbb{F}_p[t^{p-1}] \oplus \theta t^{p-2}\mathbb{F}_p[t^{p-1}] \subset H^*_{\mathbb{Z}/p}(\mathbb{F}_p),$$

$$\text{(5-10)}\qquad H^*_{\mathrm{Sym}_p}(\mathbb{F}_p(1)) = t^{(p-1)/2}\mathbb{F}_p[t^{p-1}] \oplus \theta t^{(p-3)/2}\mathbb{F}_p[t^{p-1}] \subset H^*_{\mathbb{Z}/p}(\mathbb{F}_p).$$

See eg [50, Lemma 1.4]. Let $\mathcal{C}$ be a general complex of $\mathbb{F}_p$–vector spaces, and consider Sym_p acting on its tensor power $\mathcal{C}^{\otimes p}$ by permuting the factors with Koszul signs. In this situation, there is a canonical equivariant diagonal map

$$\text{(5-11)}\qquad H^l(\mathcal{C}) \to H^{pl}_{\mathrm{Sym}_p}(\mathcal{C}^{\otimes p} \otimes \mathbb{F}_p(l))$$

which lifts the standard diagonal $H^l(\mathcal{C}) \to H^{pl}(\mathcal{C}^{\otimes p})$; see eg [50, Lemma 1.1(iv)] for its well-definedness. The equivariant diagonal is compatible with multiplication by elements of $\mathbb{F}_p$, but not additive. It is sometimes convenient to simplify the discussion of (5-11) by restricting to the cyclic subgroup:

$$\text{(5-12)}\qquad \begin{tikzcd} & H^{pl}_{\mathrm{Sym}_p}(\mathcal{C}^{\otimes p} \otimes \mathbb{F}_p(l)) \arrow[d, hook] \\ H^k(\mathcal{C}) \arrow[ur, "\text{(5-11)}"] \arrow[r] & H^{pl}_{\mathbb{Z}/p}(\mathcal{C}^{\otimes p}) \end{tikzcd}$$

By explicit computation in (5-5), one sees that the equivariant diagonal for $\mathbb{Z}/p$ becomes additive after multiplying with t. It follows that (5-11) becomes additive after multiplying with t^{p-1}, since that lies in the subgroup (5-9).

5b Cohen operations

Let $\mathcal{C}$ be a complex of $\mathbb{F}_p$–vector spaces, which has the structure of an algebra over the $\mathbb{F}_p$–coefficient version of the Fulton–MacPherson operad. Recall that the action of Sym_p on FM_p is free. The associated Cohen operation is a map

$$\mathrm{Coh}_p\colon H^l(\mathcal{C}) \to (H^*(\mathrm{FM}_p/\mathrm{Sym}_p;\mathbb{F}_p(l)) \otimes H^*(\mathcal{C}))^{pl}, \tag{5-13}$$

defined as follows:

$$\begin{tikzcd}
H^l(\mathcal{C}) \arrow[r, "(5\text{-}11)"] \arrow[ddddd, dashed] & H^{pl}_{\mathrm{Sym}_p}(\mathcal{C}^{\otimes p} \otimes \mathbb{F}_p(l)) \arrow[d, "\text{operad structure}"'] \\
& H^{pl}_{\mathrm{Sym}_p}(\mathrm{Hom}(C_{-*}(\mathrm{FM}_p), \mathcal{C}) \otimes \mathbb{F}_p(l)) \arrow[d, equal] \\
& H^{pl}_{\mathrm{Sym}_p}(\mathrm{Hom}(C_{-*}(\mathrm{FM}_p) \otimes \mathbb{F}_p(l), \mathcal{C})) \arrow[d, equal, "\text{Künneth}"'] \\
& \mathrm{Hom}^{pl}(H_{-*}^{\mathrm{Sym}_p}(\mathrm{FM}_p;\mathbb{F}_p(l)), H^*(\mathcal{C})) \arrow[dd, equal, "\text{freeness of the action}"'] \\
& \\
\left(\begin{matrix} H^*(\mathcal{C}) \\ \otimes \\ H^*(\mathrm{FM}_p/\mathrm{Sym}_p;\mathbb{F}_p(l)) \end{matrix}\right)^{pl} \arrow[r, equal] & \mathrm{Hom}^{pl}(H_{-*}(\mathrm{FM}_p/\mathrm{Sym}_p;\mathbb{F}_p(l)), H^*(\mathcal{C}))
\end{tikzcd} \tag{5-14}$$

On the middle lines, the Sym_p–action is trivial on the $\mathcal{C}$–factor. Because their definition involves (5-11), these operations are not expected to be additive. Note that we could also have defined our operations using $\mathbb{Z}/p$, but of course, they would still lie in the subspace $H^*(\mathrm{FM}_p/\mathrm{Sym}_p;\mathbb{F}_p(l)) \subset H^*(\mathrm{FM}_p/(\mathbb{Z}/p);\mathbb{F}_p)$. For computational purposes, let's spell out what happens when one decodes (5-14).

Lemma 5.1 *Suppose we have an cycle $c \in \mathcal{C}$ of degree l. Then $c^{\otimes p} \in \mathcal{C}^{\otimes p}$ is a cycle which is Sym_p–invariant up to an $\mathbb{F}_p(l)$–twist, and which therefore represents a class in*

$H^*_{\mathrm{Sym}_p}(\mathcal{C}^{\otimes p} \otimes \mathbb{F}_p(l))$. Similarly, suppose that we have a chain $B \in C_*(\mathrm{FM}_p; \mathbb{F}_p)$ with the property that ∂B goes to zero in $C_*(\mathrm{FM}_p) \otimes_{\mathrm{Sym}_p} \mathbb{F}_p(l)$. Such a chain represents a class $[B] \in H_*(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(l))$. As a consequence of the properties of B and c, the image of $B \otimes c^{\otimes p}$ under the operad action is a cycle in $\mathcal{C}$. That cycle represents the image of $[c]$ under (5-13), *paired with $[B]$.*

The structure of Cohen operations was determined in [14, Theorems 5.2 and 5.3]. The group relevant for operations on the even-degree cohomology of $\mathcal{C}$ is

$$H^*(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p) \cong \begin{cases} \mathbb{F}_p & \text{if } * = 0, 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5-15}$$

For $* = 0$, that just recovers the p–fold power for the product that is part of the Gerstenhaber algebra structure on $H^*(\mathcal{C})$. If we suppose that $p > 2$, the operation obtained from the $* = 1$ group can again be described as part of the Gerstenhaber structure, as $x \mapsto [x, x]x^{p-2}$. The twisted counterpart is more interesting:

$$H^*(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(1)) \cong \begin{cases} \mathbb{F}_p & \text{if } * = p-1,\ p-2, \\ 0 & \text{otherwise.} \end{cases} \tag{5-16}$$

Moreover, still as part of [14, Theorem 5.3], the pullback map

$$H^*_{\mathrm{Sym}_p}(\mathbb{F}_p(1)) = H^*(B\mathrm{Sym}_p; \mathbb{F}_p(1)) \to H^*(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(1)) \tag{5-17}$$

is onto. Therefore, the groups (5-16) can be thought of as generated by $t^{(p-1)/2}$ and $\theta t^{(p-3)/2}$, the lowest-degree generators in (5-10). Note that for $p > 2$, $t^{(p-1)/2}$ is the image of $\theta t^{(p-3)/2}$ under the Bockstein β.

5c Quantum Steenrod operations

The same idea works for any operad, and in particular, Deligne–Mumford spaces. Concretely, this means that we consider the action of Sym_p on DM_p which keeps the marked point z_0 fixed, and permutes $(z_1, \dots, z_p)$. Given an algebra $\mathcal{C}$ over the $\mathbb{F}_p$–coefficient Deligne–Mumford operad, one gets operations analogous to (5-13),

$$H^l(\mathcal{C}) \to \big(H^*(\mathcal{C}) \otimes H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p; \mathbb{F}_p(l))\big)^{pl}. \tag{5-18}$$

In principle, the same caveat as in Section 4f applies, which means that we should replace DM_p by a homotopy equivalent space (4-63). However, that makes no difference for the present discussion, since only the equivariant cohomology of the space will be involved.

Unfortunately, the equivariant mod p cohomology of Deligne–Mumford space is not known (to this author, at least), but there are simplified versions of this construction which are easier to understand. Because p is assumed to be prime, the Sym_p–action on DM_p has a unique orbit O_p with isotropy subgroups isomorphic to $\mathbb{Z}/p$ (all other isotropy subgroups have orders not divisible by p). For concreteness, we just look at one specific point $\diamond \in O_p$, whose isotropy subgroup is the standard cyclic subgroup $\mathbb{Z}/p \subset \mathrm{Sym}_p$:

$$\diamond = (C = \bar{\mathbb{C}} = \mathbb{C} \cup \{\infty\},\ z_0 = \infty,\ z_k = e^{2\pi\sqrt{-1}k/p} \text{ for } k = 1, \dots, p). \tag{5-19}$$

Lemma 5.2 *Restriction to $\diamond \in O_p$ yields isomorphisms*

$$\begin{tikzcd} H^*_{\mathrm{Sym}_p}(O_p; \mathbb{F}_p(l)) \arrow[r] \arrow[rr, bend right, "\cong"'] & H^*_{\mathbb{Z}/p}(O_p; \mathbb{F}_p) \arrow[r] & H^*_{\mathbb{Z}/p}(\diamond; \mathbb{F}_p) = H^*_{\mathbb{Z}/p}(\mathbb{F}_p). \end{tikzcd} \tag{5-20}$$

Proof This is elementary: the underlying map of Borel constructions, obtained by composing

$$E\mathrm{Sym}_p \times_{\mathbb{Z}/p} \diamond \hookrightarrow E\mathrm{Sym}_p \times_{\mathbb{Z}/p} O_p \twoheadrightarrow E\mathrm{Sym}_p \times_{\mathrm{Sym}_p} O_p, \tag{5-21}$$

is a homeomorphism. Moreover, the local system on $E\mathrm{Sym}_p \times_{\mathrm{Sym}_p} O_p$ associated to $\mathbb{F}_p(1)$ is canonically trivial. □

Quantum Steenrod operations are obtained by replacing DM_p in (5-18) by its subspace O_p. In view of Lemma 5.2, we can equivalently define them using the $\mathbb{Z}/p$–equivariant cohomology of a point. Written in that way, they have the form

$$Q\,\mathrm{St}_p \colon H^l(\mathcal{C}) \to (H^*(\mathcal{C}) \otimes H^*_{\mathbb{Z}/p}(\mathbb{F}_p))^{pl}. \tag{5-22}$$

Following our discussion of (5-11), we know that (5-22) becomes additive after multiplication with t^{p-1}. Since that multiplication acts injectively on $H^*_{\mathbb{Z}/p}(\mathbb{F}_p)$, one sees that (5-22) is already additive.

As an intermediate object between the two spaces considered so far, take DM_p° be the moduli space of smooth genus-zero curves with $p+1$ marked points, or equivalently (3-7), which is an open subset of DM_p containing O_p. Similarly, let FM_p° be the configuration space (3-1), which is the interior of FM_p and hence homotopy equivalent to the whole space. The forgetful map $\mathrm{FM}_p \to \mathrm{DM}_p$ restricts to a circle bundle $\mathrm{FM}_p^\circ \to \mathrm{DM}_p^\circ$.

Lemma 5.3 *Restriction to $O_p \subset \mathrm{DM}_p^\circ$, together with Lemma 5.2, yields isomorphisms*

$$H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p) \cong H^*_{\mathbb{Z}/p}(\mathrm{point};\mathbb{F}_p), \tag{5-23}$$

$$H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p(1)) \cong \begin{cases} 0 & \text{for } * < p-2, \\ H^*_{\mathbb{Z}/p}(\mathrm{point};\mathbb{F}_p) & \text{for } * \geq p-2. \end{cases} \tag{5-24}$$

Moreover, the pullback map is an isomorphism

$$H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p(1)) \to H^*_{\mathrm{Sym}}(\mathrm{FM}_p^\circ;\mathbb{F}_p(1)) \quad \text{for } * = p-2,\ p-1. \tag{5-25}$$

Proof Consider the Gysin sequence and its restriction to (5-19):

$$\begin{tikzcd} \cdots \arrow[r] & H^{*-2}_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p(l)) \arrow[r] \arrow[d] & H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p(l)) \arrow[r] \arrow[d] & H^*_{\mathrm{Sym}_p}(\mathrm{FM}_p^\circ;\mathbb{F}_p(l)) \arrow[r] \arrow[d] & \cdots \\ \cdots \arrow[r] & H^{*-2}_{\mathbb{Z}/p}(\diamond;\mathbb{F}_p) \arrow[r, "-t"] & H^*_{\mathbb{Z}/p}(\diamond;\mathbb{F}_p) \arrow[r] & H^*_{\mathbb{Z}/p}(S^1;\mathbb{F}_p) \arrow[r] & \cdots \end{tikzcd} \tag{5-26}$$

Over $\diamond$, the fiber of the circle bundle $\mathrm{FM}_p^\circ \to \mathrm{DM}_p^\circ$ can be identified with the representation of $\mathbb{Z}/p$ with weight -1. In other words, the S^1 in (5-26) carries the action of $\mathbb{Z}/p$ by clockwise rotation. The $-t$ appearing in the sequence is the associated equivariant Euler class. For $l = 0$, inspection of (5-15) shows that the rightmost $\downarrow$ in (5-26) is always an isomorphism. One can therefore prove (5-23) by upwards induction on degree.

For $l = 1$, we use a variant of the same argument. The Gysin sequence and (5-16) imply that

$$H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p(1)) \cong \begin{cases} 0 & \text{for } * < p-2, \\ H^*_{\mathrm{Sym}_p}(\mathrm{FM}_p^\circ;\mathbb{F}_p(1)) & \text{for } * = p-2. \end{cases} \tag{5-27}$$

Let's look at the first nontrivial degree, and the maps

$$H^{p-2}_{\mathrm{Sym}_p}(\mathrm{point};\mathbb{F}_p(1)) \xrightarrow{\text{pullback}} H^{p-2}_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p(1)) \xrightarrow{\text{restriction}} H^{p-2}_{\mathbb{F}_p}(\diamond;\mathbb{F}_p). \tag{5-28}$$

From (5-27) and (5-17), it follows that the first map is an isomorphism. The composition of the two maps is just (5-7), hence an isomorphism. It follows that the second map must be an isomorphism as well, which is part of (5-24). On the other hand, since the Sym_p–action has isotropy groups of order coprime to p outside O_p,

$$H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ, O_p;\mathbb{F}_p(1)) \cong H^*(\mathrm{DM}_p^\circ/\mathrm{Sym}_p, O_p/\mathrm{Sym}_p;\mathbb{F}_p(1)), \tag{5-29}$$

and the right-hand side vanishes in high degrees. From that and Lemma 5.2, one sees that the restriction map $H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ;\mathbb{F}_p(1)) \to H^*_{\mathbb{Z}/p}(\mathrm{point};\mathbb{F}_p)$ is an isomorphism in high degrees. By downward induction on degree, using (5-16) and (5-26), one obtains the degree $\geq p-1$ part of (5-24). From (5-24), it also follows that the map $H^{p-2}_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ; \mathbb{F}_p(1)) \to H^p_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ; \mathbb{F}_p(1))$ in the top row of (5-26) is an isomorphism, which then implies (5-25) in degree $p-1$; the degree $p-2$ part of the same statement has been derived before, in (5-27). $\square$

As an immediate consequence, suppose that $\mathcal{C}$ is an algebra over the chain-level Deligne–Mumford operad. Consider its induced structure as an algebra over the Fulton–MacPherson operad. By definition, the associated operations (5-13) and (5-22) fit into a commutative diagram

$$\begin{tikzcd} & \big(H^*(\mathcal{C}) \otimes H^*(\mathrm{FM}_p^\circ/\mathrm{Sym}_p; \mathbb{F}_p(l))\big)^{pl} \\ H^l(\mathcal{C}) \arrow[ur, "\text{Cohen}"] \arrow[r] \arrow[dr, "\text{quantum Steenrod}"'] & \big(H^*(\mathcal{C}) \otimes H^*_{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ; \mathbb{F}_p(l))\big)^{pl} \arrow[u] \arrow[d] \\ & \big(H^*(\mathcal{C}) \otimes H^*_{\mathbb{Z}/p}(\mathbb{F}_p)\big)^{pl} \end{tikzcd} \tag{5-30}$$

If l is odd, then Lemma 5.3 shows that both vertical arrows are isomorphisms on the degree $p-2$ or $p-1$ cohomology groups of the moduli spaces. Those cohomology groups are one-dimensional, and their generators can be identified with $\theta t^{(p-3)/2}$ and $t^{(p-1)/2}$, respectively. To put it more succinctly:

Lemma 5.4 *The Cohen and quantum Steenrod operations*

$$H^l(\mathcal{C}) \to H^{pl-k}(\mathcal{C}) \quad \text{for } l \text{ odd and } k = p-2 \text{ or } k = p-1 \tag{5-31}$$

coincide.

6 Prime power maps

This section brings together the lines of thought from Sections 4 (formal group structure) and 5 (cohomology operations). Our first task is to make part of the discussion in Section 5b more concrete, by introducing an explicit cocycle which generates $H^{p-1}(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(1))$. By looking at the relation between that cocycle and the map $\mathrm{MWW}_{1,\dots,1} \to \mathrm{FM}_p$, we obtain an abstract analogue of Theorem 1.9 in the operadic context.

6a A cocycle in unordered Fulton–MacPherson space

Take (3-5), modify it by rotation by i so as to put the resulting configurations on the imaginary axis in $\mathbb{C}$, and then compose that with projection to $\mathrm{FM}_p/\mathrm{Sym}_p$ (recall that the Sym_p–action is free, so the quotient is again a smooth manifold with corners, or topologically a manifold with boundary). The outcome is a submanifold (a copy of the associahedron S_p)

$$Z_p \subset \mathrm{FM}_p/\mathrm{Sym}_p, \tag{6-1}$$

with $\partial Z_p \subset \partial \mathrm{FM}_p/\mathrm{Sym}_p$. By definition, (6-1) has a preferred lift to FM_p, and therefore, the local system $\mathbb{F}_p(1)|Z_p$ has a canonical trivialization. Using that and the orientations of S_p and FM_p, we get a class

$$[Z_p] \in H_{p-2}(\mathrm{FM}_p/\mathrm{Sym}_p, \partial\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(1)) \cong H^{p-1}(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(1)). \tag{6-2}$$

In terms of the previous computations (5-10) and (5-16), this can be expressed as follows:

Lemma 6.1 *For* $p = 2$, (6-2) *is the image of* $\theta = t^{1/2}$ *under* (5-17); *for* $p > 2$, *it is the image of*

$$(-1)^{(p-1)/2}\Big(\frac{p-1}{2}!\Big)t^{(p-1)/2} \in H^{p-1}(B\mathrm{Sym}_p; \mathbb{F}_p(1)). \tag{6-3}$$

Proof Let's consider the more interesting case $p > 2$ first. Take the map

$$\mathrm{Conf}_p(\mathbb{C}) \to \mathbb{R}^p/\mathbb{R} = \mathbb{R}^{p-1} \tag{6-4}$$

which projects ordered configurations to their real part, and then quotients out by the diagonal $\mathbb{R}$ subspace. This map is Sym_p–equivariant, and the fiber at 0 is the subspace $\tilde{Z}_p = \mathbb{R} \times \mathrm{Conf}_p(\mathbb{R}) \subset \mathrm{Conf}_p(\mathbb{C})$ of configurations with common real part, $(z_1 = s + \sqrt{-1}\, t_1, \dots, z_p = s + \sqrt{-1}\, t_p)$. Let's orient that by using the coordinates $(s, t_1, \dots, t_p)$ in this order. This differs from its orientation as a fiber of (6-4) by a Koszul sign $(-1)^{p(p-1)/2} = (-1)^{(p-1)/2}$. On the other hand, the fiber at 0 represents the pullback via (6-4) of the equivariant Euler class of the $\mathbb{Z}/p$–representation $\mathbb{R}^p/\mathbb{R}$. From this and (10-5), we get

$$[\tilde{Z}_p] = (-1)^{(p-1)/2}\Big(\frac{p-1}{2}!\Big)t^{(p-1)/2} \in H^{p-1}_{\mathbb{Z}/p}(\mathrm{Conf}_p(\mathbb{C}); \mathbb{F}_p). \tag{6-5}$$

The corresponding relation must hold in $H^{p-1}_{\mathrm{Sym}_p}(\mathrm{Conf}_p(\mathbb{C}); \mathbb{F}_p(1))$ as well, since both classes involved live in that group, and the map from there to $\mathbb{Z}/p$–equivariant cohomology is injective. Note that $\tilde{Z}_p$ is the preimage of Z_p under the quotient map

$\mathrm{Conf}_p(\mathbb{C}) \to \mathrm{FM}_p$. Moreover, inspection of (3-6) shows that the orientations of FM_p, Z_p and $\widetilde{Z}_p$ we have used are compatible with that relation. Since the quotient map is equivariant and a homotopy equivalence, (6-5) implies the corresponding property for $[Z_p]$.

One could follow the same strategy for $p = 2$, but we can be even more explicit. The generator of $H_1(\mathrm{FM}_2/\mathrm{Sym}_2; \mathbb{Z}/2) \cong \mathbb{Z}/2$ consists of a loop of configurations where two points rotate around each other, and its image in $H_1(B\mathrm{Sym}_2; \mathbb{Z}/2) \cong \mathbb{Z}/2$ is obviously nontrivial. On the other hand, that loop intersects Z_p transversally at exactly one point, which proves the desired statement. $\square$

6b A cycle in unordered Fulton–MacPherson space

Consider the space $\mathrm{MWW}_{1,\dots,1}$ with d colors, denoted here by $\mathrm{MWW}_{\underline{d}}$ for the sake of brevity. As a special case of (3-11), its codimension-one boundary faces are of the form

$$\mathrm{MWW}_{\underline{d_1}} \times \cdots \times \mathrm{MWW}_{\underline{d_r}} \times S_r \xrightarrow{T_{I_1,\dots,I_r}} \mathrm{MWW}_{\underline{d}}; \tag{6-6}$$

there is one such face for each decomposition of $\{1, \dots, d\}$ into $r \geq 2$ nonempty subsets $(I_1, \dots, I_r)$ with $d_k = |I_k|$. In describing the boundary faces, we have used the identifications (3-18). Suppose that we choose maps (3-13) so as to be compatible with (3-18), as in Section 4e. Consider two decompositions $(I_1, \dots, I_d)$ and $(\tilde{I}_1, \dots, \tilde{I}_d)$, which correspond to the same ordered partition $d_k = |I_k| = |\tilde{I}_k|$. Then, the associated maps (6-6) and (3-13) fit into a commutative diagram

$$\begin{tikzcd}
\mathrm{MWW}_{\underline{d}} \arrow[d] & \mathrm{MWW}_{\underline{d_1}} \times \cdots \times \mathrm{MWW}_{\underline{d_r}} \times S_r \arrow[l, "T_{I_1,\dots,I_r}"'] \arrow[r, "T_{\tilde{I}_1,\dots,\tilde{I}_r}"] & \mathrm{MWW}_{\underline{d}} \arrow[d] \\
\mathrm{FM}_d & \mathrm{FM}_d \arrow[l, "\sigma_{I_1,\dots,I_r}"', "\cong"] \arrow[r, "\sigma_{\tilde{I}_1,\dots,\tilde{I}_r}", "\cong"'] & \mathrm{FM}_d
\end{tikzcd} \tag{6-7}$$

Here, $\sigma_{I_1,\dots,I_r} \in \mathrm{Sym}_d$ is the unique permutation which maps $\{1, \dots, d_1\}$ order-preservingly to I_1, $\{d_1 + 1, \dots, d_1 + d_2\}$ order-preservingly to I_2, and so on; and correspondingly for $\sigma_{\tilde{I}_1,\dots,\tilde{I}_r}$. Suppose that we choose fundamental chains to be also compatible with (3-18). Then, (4-24) simplifies to

$$\partial[\mathrm{MWW}_{\underline{d}}] = \sum_{(I_1,\dots,I_r)} \pm T_{I_1,\dots,I_r,*}([\mathrm{MWW}_{\underline{d_1}}] \times \cdots \times [\mathrm{MWW}_{\underline{d_r}}] \times [S_r]). \tag{6-8}$$

Thinking of the homology of $\mathrm{FM}_p/\mathrm{Sym}_p$ as in Lemma 5.1, we get:

Lemma 6.2 *Suppose that $d = p$ is prime. Then, the image of* $[\mathrm{MWW}_{\underline{p}}]$ *under* (3-13) *is a chain, denoted here by $B_p \in C_*(\mathrm{FM}_p; \mathbb{F}_p)$, with the property that ∂B_p goes to zero in $C_*(\mathrm{FM}_p; \mathbb{F}_p) \otimes_{\mathrm{Sym}_p} \mathbb{F}_p(1)$. Therefore, it represents a class $[B_p]$ in $H_{p-1}(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(1))$.*

Proof Consider two codimension-one boundary faces as in (6-7). The resulting chains in FM_d differ by applying the permutation $\sigma_{\tilde{I}_1,\dots,\tilde{I}_r}\sigma^{-1}_{I_1,\dots,I_r}$. Hence, when mapped to $C_*(\mathrm{FM}_d) \otimes_{\mathrm{Sym}_p} \mathbb{F}_p(1)$, they differ by the sign of that permutation. On the other hand, their entries in (6-8) differ by the same sign. When computing ∂B_p in $C_*(\mathrm{FM}_p) \otimes_{\mathrm{Sym}_p} \mathbb{F}_p(1)$, the two kinds of signs cancel, which means that the contributions are the same. Now, the cyclic group $\mathbb{Z}/p \subset \mathrm{Sym}_p$ acts freely on ordered decompositions corresponding to the same ordered partition, and this provides the required cancellation mod p for the terms of ∂B_p. □

Lemma 6.3 *The canonical pairing between the cohomology class of* (6-2) *and the homology class from Lemma 6.2 is $[Z_p] \cdot [B_p] = (-1)^{p(p-1)/2}$.*

Proof We think of this Poincaré-dually as an intersection number. The relevant cycles intersect at exactly one point of FM_p, which is a configuration $(z_1; \dots; z_p)$ with $\mathrm{re}(z_1) = \dots = \mathrm{re}(z_d) = 0$ and $\mathrm{im}(z_1) < \dots < \mathrm{im}(z_d)$. The tangent space of $Z_d \subset \mathrm{FM}_d$ at that point can be thought of as keeping (z_1, z_2) fixed, and moving $(z_3, \dots, z_d)$ infinitesimally in the imaginary direction. The tangent space to the image of $\mathrm{MWW}_{\underline{p}}$ at the same point consists of keeping z_1 fixed, but moving $(z_2, \dots, z_d)$ infinitesimally in the real direction. Note that positive horizontal motion of z_2 yields a clockwise motion of the angular component of $z_2 - z_1$. This observation, when combined with standard Koszul signs, yields the desired local intersection number. □

To see the implications of Lemmas 6.1 and 6.3, note that, by the dual of (5-16) and (5-17),

$$H_{p-1}(\mathrm{FM}_p/\mathrm{Sym}_p; \mathbb{F}_p(1)) \cong H_{p-1}(B\mathrm{Sym}_p; \mathbb{F}_p(1)) \cong \mathbb{F}_p. \tag{6-9}$$

In those terms, what we have shown is:

Lemma 6.4 *For $p = 2$, the homology class of the cycle from Lemma 6.2 is the unique nontrivial element in* (6-9)*. For $p > 2$, it is*

$$\Big(\frac{p-1}{2}!\Big)^{-1}$$

times the standard generator (dual to $t^{(p-1)/2}$).

Take $[c] \in H^{\mathrm{odd}}(\mathcal{C})$ and our B_p, and consider the image of $B_p \otimes c^{\otimes p}$ under the operad action. This defines a map $H^{\mathrm{odd}}(\mathcal{C}) \to H^{\mathrm{odd}}(\mathcal{C})$, which by Lemma 5.1 is a certain component of the Cohen operation applied to $[c]$. Lemma 6.4 tells us exactly what it is:

$$\begin{cases} \text{the } t^{1/2} \text{ (or } \theta\text{) component of (5-14)} & \text{if } p = 2, \\ \left(\frac{p-1}{2}!\right)^{-1} \text{ times the } t^{(p-1)/2}\text{–component of (5-14)} & \text{if } p > 2. \end{cases} \tag{6-10}$$

If c has degree 1, the same process computes $\beta_{\mathcal{C}}^{1,\dots,1}(c;\dots;c)$, by definition of (4-26). Lemma 4.18 shows that this is the leading order term of $\Pi_{\mathcal{C}}^p(\gamma,\dots,\gamma)$ for a Maurer–Cartan element $\gamma = qc + O(q^2)$, and Corollary 4.24 identifies that with the p–fold product of γ under our formal group law. The consequence, under the assumption of homological unitality inherited from the proof of Proposition 4.20, is:

Theorem 6.5 *Take the group law* $\bullet$ *on* $\mathrm{MC}(\mathcal{C}; q\mathbb{F}_p[q]/q^{p+1})$*. The* p^{th} *power map for that group fits into a commutative diagram like* (1-13), *with the operation* (6-10) *at the bottom.*

To conclude our discussion, note that if the operad structure is induced from one over Deligne–Mumford spaces, as in (4-64), then the relevant operation (6-10) can also be written as a quantum Steenrod operation, by Lemma 5.4.

7 Constructions using pseudoholomorphic curves

We will now translate the previous arguments into more specifically symplectic terms. The choice of singular chains on parameter spaces, and its application to a general operadic structure, is replaced by a choice of perturbations which make the moduli spaces regular, followed by counting-of-solutions to extract the algebraic operations. For technical convenience, we use Hamiltonian Floer theory (with a small time-independent Hamiltonian) as a model for cochains on our symplectic manifold. Correspondingly, all the operations are defined using inhomogeneous Cauchy–Riemann equations on punctured surfaces. This makes no difference with respect to Theorem 1.9, since the Floer-theoretic version of quantum Steenrod operations agrees with that defined using ordinary pseudoholomorphic curves (for $p = 2$, see [75]; the same strategy works for general p).

7a Floer-theoretic setup

Let (X, ω_X) be a closed symplectic manifold, which is monotone (1-3). We fix a C^2–small Morse function $H \in C^\infty(X, \mathbb{R})$, with its Hamiltonian vector field Z_H. We also fix a compatible almost complex structure J, and the associated metric g_J. We require:

Properties 7.1 (i) All spaces of Morse flow lines for (H, g_J) are regular.

(ii) All one-periodic orbits of Z_H are constant, hence critical points of H. Moreover, the linearized flow at each such point x is nondegenerate for all times $t \in (0, 1]$, which implies that the Conley–Zehnder index is equal to the Morse index $\mu(x)$.

(iii) No J–holomorphic sphere v with $c_1(v) = 1$ passes through a critical point of H or intersects an isolated Morse flow line. Here, $c_1(v)$ is the usual shorthand for $c_1(X)$ integrated over $[v] \in H_2(X)$.

Consider the autonomous Floer equation, where as usual $S^1 = \mathbb{R}/\mathbb{Z}$:

$$\begin{cases} u\colon \mathbb{R} \times S^1 \to X, \\ \partial_s u + J(\partial_t u - Z_H) = 0, \\ \lim_{s \to \pm\infty} u(s,t) = x_\pm, \end{cases} \tag{7-1}$$

where the limits $x_\pm$ are critical points of H. This equation has an $(\mathbb{R}\times S^1)$–symmetry by translation in both directions. For S^1–invariant solutions, meaning t–independent maps $u = u(s)$, it reduces to the negative gradient flow equation $du/ds + \nabla_{g_J} H = 0$. Denote by D_u the linearized operator at a solution of (7-1). Its Fredholm index can be computed as

$$\operatorname{index}(D_u) = \mu(x_-) - \mu(x_+) + 2c_1(u), \tag{7-2}$$

where in the last term, we have extended u to $(\mathbb{R} \times S^1) \cup \{\pm\infty\} = S^2$, the two-point compactification. As a consequence of transversality results in [33; 20] (see in particular [33, Theorem 7.3] and [20, Theorem 7.4]), we may further require:

Properties 7.2 (i) All solutions of Floer's equation with $\operatorname{index}(D_u) \leq 1$ are independent of t. Concretely, there are none with negative index; the only ones with index zero are constant; and those with $\operatorname{index}(D_u) = 1$ are isolated Morse flow lines for (H, g_J).

(ii) For the last-mentioned u, all solutions of $D_u \xi = 0$ are independent of t, hence lie in the kernel of the corresponding linearized operator from Morse theory. In view of Properties 7.1(i) and (ii), this implies that such Floer solutions are regular.

Define $\mathcal{C}$ to be the standard Morse complex for (H, g_J), meaning that

$$\mathcal{C} = \bigoplus_x \mathbb{Z}_x[-\mu(x)], \tag{7-3}$$

where $\mathbb{Z}_x$ is the orientation line (the rank-one free abelian group whose two generators correspond to orientations of the descending manifold of x), with a differential $d_{\mathcal{C}}$ that counts isolated gradient flow lines. When considered as a $\mathbb{Z}/2$–graded space, this is equal to the Floer complex of (H, J), thanks to the properties above. Our conventions are cohomological, meaning that with notation as in (7-1), $d_{\mathcal{C}}$ takes "x_+ to x_-".

7b Operations

Take $\overline{\mathbb{C}} = \mathbb{C}P^1$ with marked points $z_0 = \infty$ and $z_1, \dots, z_d \in \mathbb{C}$. Consider the resulting punctured surface,

$$C = \overline{\mathbb{C}} \setminus \{z_0, \dots, z_d\} = \mathbb{C} \setminus \{z_1, \dots, z_d\}. \tag{7-4}$$

An inhomogeneous term ν_C is a $(0,1)$–form on C with values on vector fields on X:

$$\nu_C \in \Omega^{0,1}(C, C^\infty(X, TX)) = C^\infty(C \times X, \mathrm{Hom}_{\mathbb{C}}(\overline{TC}, TX)), \tag{7-5}$$

where the $(0,1)$ part is taken with respect to J. We require that our inhomogeneous terms should have a special structure near the marked points:

$$\nu_C = \begin{cases} \big(Z_H \otimes \mathrm{re}(d\log(z - z_k)/2\pi\sqrt{-1})\big)^{0,1} & \text{near } z_k \text{ for } k > 0, \\ \big(Z_H \otimes \mathrm{re}(d\log(z - \xi_C)/2\pi\sqrt{-1})\big)^{0,1} & \text{near } z_0 = \infty, \end{cases} \tag{7-6}$$

where $\xi_C \in \mathbb{C}$ is an auxiliary datum that we consider as part of ν_C. In cylindrical coordinates

$$z = \begin{cases} z_k + \exp(-2\pi(s + \sqrt{-1}t)) & \text{near } z_k \text{ for } k > 0, \text{ where } (s,t) \in \mathbb{R}^{\geq 0} \times S^1, \\ \xi_C + \exp(-2\pi(s + \sqrt{-1}t)) & \text{near } z_0 = \infty, \text{ where } (s,t) \in \mathbb{R}^{\leq 0} \times S^1, \end{cases} \tag{7-7}$$

what (7-6) says is $\nu_C = (Z_H \otimes dt)^{0,1}$. Consider the inhomogeneous Cauchy–Riemann equation

$$\begin{cases} u \colon C \to X, \\ \bar\partial u = \nu_C(u), \\ \lim_{z \to z_k} u(z) = x_k, \end{cases} \tag{7-8}$$

where the limits x_k are again critical points of H. When written in coordinates (7-7) near the marked points, (7-8) reduces to (7-1), explaining why this convergence

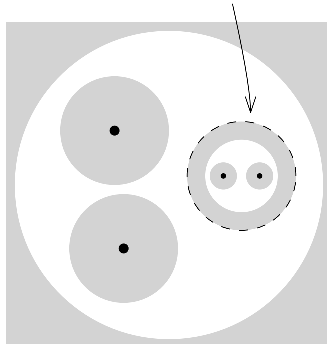

Figure 15: Gluing punctured planes equipped with inhomogeneous terms, as in (7-12). The shaded regions are where the inhomogeneous term on the glued surface is prescribed.

condition makes sense. The linearization of (7-8) has

$$\operatorname{index}(D_u) = \mu(x_0) - \sum_{j=1}^{d} \mu(x_j) + 2c_1(u). \tag{7-9}$$

We will not explain the compactness and transversality theory for moduli spaces of solutions of (7-8), both being standard (the first due to monotonicity, the second because we have complete freedom in choosing ν_C on a compact part of C). For a single surface C and a generic choice of ν_C, counting solutions of (7-8) will give rise to a chain map $\mathcal{C}^{\otimes d} \to \mathcal{C}$ which preserves the $\mathbb{Z}/2$–degree (and which represents the d–fold pair-of-pants product).

We need to review briefly the gluing process for surfaces, to see how it fits in with inhomogeneous terms. Suppose that we have two surfaces $C_k = \mathbb{C} \setminus \{z_{k,1}, \dots, z_{k,d_k}\}$ for $k = 1, 2$, which also come with inhomogeneous terms ν_{C_k}, and in particular $\xi_{C_k} \in \mathbb{C}$. Fix some $0 \le i < d_1$. The gluing process produces a family of surfaces $C_\lambda = \mathbb{C} \setminus \{z_{\lambda,1}, \dots, z_{\lambda,d}\}$, where $d = d_1 + d_2 - 1$, depending on a sufficiently small parameter $\lambda > 0$. Namely, take the affine transformation

$$\phi_\lambda(z) = \lambda(z - \xi_{C_2}) + z_{1,i+1}; \tag{7-10}$$

then

$$z_{\lambda,k} = \begin{cases} z_{1,k} & \text{for } k \le i, \\ \phi_\lambda(z_{2,k-i}) & \text{for } i < k \le i + d_2, \\ z_{1,k-d_2+1} & \text{for } k > i + d_2. \end{cases} \tag{7-11}$$

We want to equip the glued surfaces with inhomogeneous terms ν_{C_λ} which are smoothly dependent on λ and have the following property. Fix some sufficiently small $r > 0$ and large $R > 0$. First,

$$\text{(7-12)}\qquad \begin{cases} \nu_{C_\lambda} = \nu_{C_1} & \text{where } |z| \geq R \text{ and } |z - z_{1,k}| \leq r \text{ for } 0 < k \neq i+1, \\ \nu_{C_\lambda} = \phi_{\lambda,*}\nu_{C_2} & \text{where } \lambda R \leq |z - z_{1,i+1}| \leq r, \\ \phi_\lambda^* \nu_{C_\lambda} = \nu_{C_2} & \text{where } |z - z_{2,k}| \leq r \text{ for any } k > 0. \end{cases}$$

The upshot is that ν_{C_λ} is completely prescribed in certain (partly λ–dependent) neighborhoods of the marked points on C_λ, as well as on an annular "gluing region"; see Figure 15. On each such region, ν_{C_λ} is given by a similar expression as in (7-6). In particular, the middle line of (7-12) really says that

$$\text{(7-13)}\qquad \nu_{C_\lambda} = \left(Z_H \otimes \mathrm{re}\left(d\frac{\log(z - z_{1,i+1})}{2\pi\sqrt{-1}} \right)\right)^{0,1} \quad \text{where } \lambda R \leq |z - z_{1,i+1}| \leq r$$

$$\iff \phi_\lambda^* \nu_{C_\lambda} = \left(Z_H \otimes \mathrm{re}\left(d\frac{\log(z - \xi_{C_2})}{2\pi\sqrt{-1}} \right)\right)^{0,1} \quad \text{where } R \leq |z - \xi_{C_2}| \leq \lambda^{-1} r.$$

Additionally, there are asymptotic conditions as $\lambda \to 0$:

(7-14)
- On $|z - z_{1,i+1}| \geq r$, the family ν_{C_λ} can be smoothly extended to $\lambda = 0$, by setting that extension equal to ν_{C_1}.
- On $|z| \leq R$, the family $\phi_\lambda^* \nu_{C_\lambda}$ can be smoothly extended to $\lambda = 0$, by setting that extension equal to ν_{C_2}.

Given that, it makes sense for a sequence of solutions of (7-8) on C_{λ_k}, with $\lambda_k \to 0$, to converge to a "broken solution" which consists of corresponding solutions on C_1 and C_2; and conversely, the gluing process for broken solutions applies — as used, for instance, in proving associativity of the pair-of-pants product. Again, we omit the details, which are standard. Thanks to our use of an autonomous Hamiltonian, there is also a version where C_2 is rotated before being glued in, meaning that we use a small $\lambda \in \mathbb{C}^*$ (inserting absolute values wherever the size of λ appears in the formulae above).

A process such as (7-11), in which the λ–dependence of the marked points follows a specific pattern, is simple to describe, but far more rigid than the analytic arguments require. Here is a more appropriate formulation, where the first part describes the ingredient for compactness arguments, and the second part addresses gluing of solutions.

Definition 7.3 Take C_1 and C_2 as before. Choose arbitrary families of surfaces with inhomogeneous terms $C_{1,r}$ and $C_{2,r}$, depending on $r \in \mathbb{R}^m$ for some m, and which reduce to the given ones for $r=0$. Apply the previously described notion of gluing in a parametrized way, which means that we have a family $C_{\lambda,r}$.

(i) Suppose that C_k is a sequence of surfaces with inhomogeneous terms ν_{C_k}, which for $k \gg 0$ are isomorphic to C_{λ_k,r_k} for $\lambda_k > 0$ and $(\lambda_k, r_k) \to (0,0)$. We then say that the C_k degenerate to (C_1, C_2).

(ii) Suppose that C_σ is a smooth family of surfaces with inhomogeneous terms ν_{C_σ}, depending on a parameter $\sigma > 0$. Suppose that for small σ, these are isomorphic to $C_{\lambda(\sigma),r(\sigma)}$, where $(\lambda(\sigma), r(\sigma))$ satisfies $(\lambda(0), r(0)) = (0,0)$ and $\lambda'(0) > 0$. We then say that the family C_σ is obtained by smoothing (C_1, C_2).

To clarify the notation, it might be useful to look slightly ahead to our first application. When defining an A_∞–structure, one deals with C_1 and C_2 which depend, respectively, on moduli in $S_{d_1} \setminus \partial S_{d_1}$ and $S_{d_2} \setminus \partial S_{d_2}$. In these terms, r is a local coordinate on the product of those spaces, while λ is the transverse coordinate to $S_{d_1} \times S_{d_2} \subset \partial S_d$, where $d = d_1 + d_2 - 1$. As this example shows, our discussion has been limited to the simplest process of gluing two surfaces together; a complete description would include the generalization to arbitrarily many surfaces.

To round off the discussion of inhomogeneous terms, let's mention an obvious generalization, which is to equip (7-4) with a family of compatible almost complex structures $(J_z)_{z\in C}$ which reduce to the given J outside a compact subset. When defining the associated notion of inhomogeneous term, one uses those structures to define the $(0,1)$ part, and similarly for (7-8). It is straightforward to extend the gluing process to this situation. Usually, this generalization is not required, since the freedom to choose ν_C is already enough to achieve transversality of moduli spaces. However, there are situations such as the construction of continuation maps, where varying almost complex structures necessarily occur (because one is trying to relate different choices of J).

7c The quantum A_∞–structure

This is the most familiar application. Given $(s_1, \dots, s_d)$ as in (3-1), we think of them as complex points $z_k = s_k$, and then equip the resulting surface (7-4) with an inhomogeneous term ν_C, which should vary smoothly in dependence on the points, and be invariant under the symmetries in (3-1); one can think of this as a fiberwise inhomogeneous term on the universal family of surfaces over $S_d \setminus \partial S_d$. Along the boundary of the moduli space, we want the family to extend along the lines indicated in Definition 7.3(ii). Of course, on a boundary stratum of codimension k, one has k components that are being glued together, and the definition should be adapted accordingly. The outcome is a parametrized moduli space, which consist of points of $S_d \setminus \partial S_d$ together with a solution of (7-8) on the associated surface. For generic choices, these parametrized moduli spaces are regular. Moreover, they are oriented relative to the orientation spaces at limit points, meaning that a choice of isomorphism $\mathbb{Z}_{x_k} \cong \mathbb{Z}$ for $k = 0, \dots, d$, determines an orientation of the parametrized moduli spaces. A signed count of points in the zero-dimensional moduli spaces, with auxiliary signs as in (4-3), yields operations $\mu_{\mathcal{C}}^d$ for $d \geq 2$, which one combines with the Floer differential $\mu_{\mathcal{C}}^1 = -d_{\mathcal{C}}$ to form the ($\mathbb{Z}/2$–graded) quantum A_∞–structure.

One can adapt the arguments from Section 4b to show that the quantum A_∞–structure is, in a suitable homotopical sense, independent of all choices, including the Floer differential. Suppose that we have (H, J) and $(\widetilde{H}, \widetilde{J})$, leading to chain complexes $(\mathcal{C}, d_{\mathcal{C}})$ and $(\widetilde{\mathcal{C}}, d_{\widetilde{\mathcal{C}}})$. For each of the two, we make the choices of inhomogeneous terms required to build the A_∞–structures, denoted by $\mu_{\mathcal{C}}$ and $\mu_{\widetilde{\mathcal{C}}}$. To relate them, we start by picking a third version of the chain complex, denoted by $(\breve{\mathcal{C}}, d_{\breve{\mathcal{C}}})$, based on some $(\breve{H}, \breve{J})$. Next, we introduce maps

$$\phi_{\breve{\mathcal{C}}}^{p,1,q} \colon \mathcal{C}^{\otimes p} \otimes \breve{\mathcal{C}} \otimes \widetilde{\mathcal{C}}^{\otimes q} \to \breve{\mathcal{C}}[1-p-q], \tag{7-15}$$

with $\phi_{\breve{\mathcal{C}}}^{0,1,0} = -d_{\breve{\mathcal{C}}}$, which turn $\breve{\mathcal{C}}[1]$ into an A_∞–bimodule, with $\mu_{\mathcal{C}}$ acting on the left and $\mu_{\widetilde{\mathcal{C}}}$ on the right. This is analogous to (4-9), except that the conditions on $\phi^{p,1,0}$ and $\phi^{0,1,q}$ which we imposed there are no longer satisfied. The geometric construction of (7-15) involves another family of inhomogeneous terms over $S_d \setminus \partial S_d$, $d = p+1+q$. Those terms are modeled on H near $(z_1, \dots, z_p)$, on $\widetilde{H}$ near $(z_{p+2}, \dots, z_d)$, and on $\breve{H}$ near the remaining points (z_0, z_{p+1}). Similarly, our surfaces carry varying families of complex structures. The behavior under degeneration to ∂S_d follows the pattern from Figure 12, with some components of the limit carrying the inhomogeneous terms that define the two A_∞–ring structures, and others, the A_∞–bimodule structure; we have represented this in a more geometric way in Figure 16.

At this point, we add continuation maps to the mix. These arise from the configuration $(z_0 = \infty,\ z_1 = 0)$, meaning the surface $C = \mathbb{C}^*$. In our application, the behavior at z_0 is always given by $(\breve{H}, \breve{J})$, and that at z_1 by either (H, J) or $(\widetilde{H}, \widetilde{J})$. The outcome is two chain maps

$$\mathcal{C} \xrightarrow{\psi_{\breve{\mathcal{C}}}^{1,0}} \breve{\mathcal{C}} \xleftarrow{\psi_{\breve{\mathcal{C}}}^{0,1}} \widetilde{\mathcal{C}}. \tag{7-16}$$

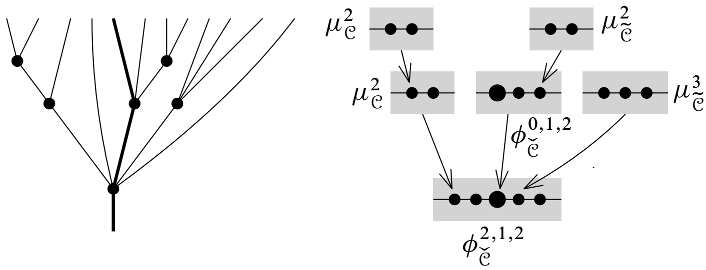

Figure 16: The A_∞–bimodule operations (7-15). We show the behavior of the inhomogeneous terms on a sample (codimension 5) boundary stratum, for $(p,q)=(4,7)$.

Our sign conventions are nonstandard: on cohomology, $\psi^{0,1}$ induces the canonical isomorphism between Floer cohomology groups, whereas $\psi^{1,0}$ has the opposite sign. Extending (7-16), we want to build operations

$$\psi^{p,q}_{\check{\mathcal{C}}}\colon \mathcal{C}^{\otimes p}\otimes\widetilde{\mathcal{C}}^{\otimes q}\to\check{\mathcal{C}}[1-p-q] \quad \text{for } d=p+q>0, \tag{7-17}$$

which unlike their counterparts in (4-12) are defined even if p or q are zero, and (that being taken into account) satisfy the same kind of relation (4-13). Geometrically, the parameter space underlying (7-17) is no longer $[0,1]\times S_d$ as in Section 4b, but instead S_{d+1}, where we think of having inserted an additional point $s_\dagger$, with $s_p<s_\dagger<s_{p+1}$, into (3-1). The orientation is that associated to ordered configurations $(s_1,\dots,s_p,s_\dagger,\dots,s_{p+q})$ multiplied by $(-1)^p$. When forming the associated surface (7-4), we do not equip it with a puncture corresponding to $s_\dagger$: the position of that point just serves as an additional modular variable. The inhomogeneous terms and almost complex structures are determined by (H,J) near $z_1,\dots,z_p$; by $(\widetilde{H},\widetilde{J})$ near $z_{p+1},\dots,z_{p+q}$; and by $\check{H}$ near $\tilde{z}_0$. In the limit as we approach a point of ∂S_{d+1}, the screen containing $s_\dagger$ corresponds to a component surface which carries data underlying (7-17), while the other components have data underlying the A_∞–ring structures or the A_∞–bimodule structure.

Example 7.4 Consider the cases where $p+q=2$. The algebraic relations are

$$\left.\begin{aligned}
\psi^{1,0}_{\check{\mathcal{C}}}(\mu^2_{\mathcal{C}}(c_1,c_2)) - \phi^{1,1,0}_{\check{\mathcal{C}}}(c_1;\psi^{1,0}_{\check{\mathcal{C}}}(c_2))\\
-\phi^{0,1,1}_{\check{\mathcal{C}}}(\psi^{1,0}_{\check{\mathcal{C}}}(c_1);\tilde{c}_2) - \phi^{1,1,0}_{\check{\mathcal{C}}}(c_1;\psi^{0,1}_{\check{\mathcal{C}}}(\tilde{c}_2))\\
\psi^{0,1}_{\check{\mathcal{C}}}(\mu^2_{\widetilde{\mathcal{C}}}(\tilde{c}_1,\tilde{c}_2)) - \phi^{0,1,1}_{\check{\mathcal{C}}}(\psi^{0,1}_{\check{\mathcal{C}}}(\tilde{c}_1);\tilde{c}_2)
\end{aligned}\right\} = \text{(terms involving differentials)}. \tag{7-18}$$

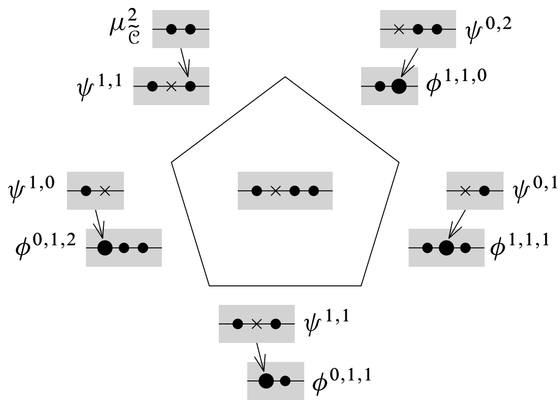

Figure 17: The construction of the maps (7-17), in the case from Example 7.5. The cross marks the additional point (not a puncture of the associated surface).

On the cohomology level, consequence is that if the classes $[c_k]$ and $[\tilde{c}_k]$ correspond to each other under canonical isomorphisms, meaning that $[\psi^{1,0}_{\check{\mathcal{C}}}(c_k)] = -[\psi^{0,1}_{\check{\mathcal{C}}}(\tilde{c}_k)]$, then their products inherit the same property:

$$
\begin{aligned}
[\psi^{1,0}_{\check{\mathcal{C}}}(\mu^2_{\mathcal{C}}(c_1, c_2))] &= [\phi^{1,1,0}_{\check{\mathcal{C}}}(c_1; \psi^{1,0}_{\check{\mathcal{C}}}(c_2))] = -[\phi^{1,1,0}_{\check{\mathcal{C}}}(c_1; \psi^{0,1}_{\check{\mathcal{C}}}(\tilde{c}_2))] \\
&= [\phi^{0,1,1}_{\check{\mathcal{C}}}(\psi^{1,0}_{\check{\mathcal{C}}}(c_1); \tilde{c}_2)] = -[\phi^{0,1,1}_{\check{\mathcal{C}}}(\psi^{0,1}_{\check{\mathcal{C}}}(\tilde{c}_1); \tilde{c}_2)] \\
&= -[\psi^{0,1}_{\check{\mathcal{C}}}(\mu^2_{\tilde{\mathcal{C}}}(\tilde{c}_1, \tilde{c}_2))].
\end{aligned}
\tag{7-19}
$$

Example 7.5 For $(p,q) = (1,2)$, the algebraic relation is

$$
\begin{aligned}
-\phi^{0,1,2}_{\check{\mathcal{C}}}(\psi^{1,0}_{\check{\mathcal{C}}}(c_1); \tilde{c}_2, \tilde{c}_3) - \phi^{0,1,1}_{\check{\mathcal{C}}}(\psi^{1,1}_{\check{\mathcal{C}}}(c_1; \tilde{c}_2); \tilde{c}_3) - \phi^{1,1,1}_{\check{\mathcal{C}}}(c_1; \psi^{0,1}_{\check{\mathcal{C}}}(\tilde{c}_2); \tilde{c}_3) \\
- \phi^{1,1,0}_{\check{\mathcal{C}}}(c_1; \psi^{0,2}_{\check{\mathcal{C}}}(\tilde{c}_2, \tilde{c}_3)) + (-1)^{\|c_1\|} \psi^{1,1}_{\check{\mathcal{C}}}(c_1; \mu^2_{\tilde{\mathcal{C}}}(\tilde{c}_2, \tilde{c}_3)) \\
= (\text{terms involving differentials}).
\end{aligned}
\tag{7-20}
$$

Figure 17 shows the relevant degenerations, corresponding to the boundary faces of S_4.

Using (7-15) and (7-17), we define an A_∞–structure on

$$
\mathcal{H} = (\mathcal{C} \otimes \mathbb{Z}u) \oplus (\tilde{\mathcal{C}} \otimes \mathbb{Z}\tilde{u}) \oplus (\check{\mathcal{C}} \otimes \mathbb{Z}v),
\tag{7-21}
$$

where the symbols u, $\tilde{u}$ and v have degrees as in (2-3). The definition is a modified version of (4-20). The differential is

$$
\begin{aligned}
\mu^1_{\mathcal{H}}(c \otimes u) &= \mu^1_{\mathcal{C}}(c) \otimes u + (-1)^{\|c\|} \psi^{1,0}_{\check{\mathcal{C}}}(c) \otimes v, \\
\mu^1_{\mathcal{H}}(\tilde{c} \otimes \tilde{u}) &= \mu^1_{\tilde{\mathcal{C}}}(\tilde{c}) \otimes \tilde{u} + (-1)^{\|\tilde{c}\|} \psi^{0,1}_{\check{\mathcal{C}}}(\tilde{c}) \otimes v, \\
\mu^1_{\mathcal{H}}(\check{c} \otimes v) &= \phi^{0,1,0}_{\check{\mathcal{C}}}(\check{c}) \otimes v,
\end{aligned}
\tag{7-22}
$$

and we similarly change the higher A_∞–operations in the case when the input consists of only terms from either $\mathcal{C}$ or $\widetilde{\mathcal{C}}$:

$$\begin{aligned}\mu^d_{\mathcal{H}}(c_1\otimes u,\dots,c_d\otimes u) &= \mu^d_{\mathcal{C}}(c_1,\dots,c_d)\otimes u+(-1)^{\maltese_d}\psi^{d,0}_{\widetilde{\mathcal{C}}}(c_1,\dots,c_d)\otimes v,\\ \mu^d_{\mathcal{H}}(\tilde{c}_1\otimes\tilde{u},\dots,\tilde{c}_d\otimes\tilde{u}) &= \mu^d_{\widetilde{\mathcal{C}}}(\tilde{c}_1,\dots,\tilde{c}_d)\otimes\tilde{u}+(-1)^{\maltese_d}\psi^{0,d}_{\widetilde{\mathcal{C}}}(\tilde{c}_1,\dots,\tilde{c}_d)\otimes v.\end{aligned} \tag{7-23}$$

As before, the projection maps from (7-21) to $\mathcal{C}$ or $\widetilde{\mathcal{C}}$ are compatible with A_∞–ring structures, and are chain homotopy equivalences. This implies the desired well-definedness statement for the A_∞–structure, as in (4-23).

7d An alternative strategy for proving independence

The approach to well-definedness of the quantum A_∞–structure adopted above involves additional families of Riemann surfaces, leading to the larger A_∞–ring $\mathcal{H}$, which serves as an intermediate object. Alternatively, as we will now explain, one can enlarge the target symplectic manifold.

Let's start by looking at a toy model, namely the symplectic manifold S^2.

(7-24) Choose (H_{S^2}, J_{S^2}) as in Section 7a, satisfying the following additional technical condition. At a local maximum or minimum of H_{S^2}, the Hessian is J_{S^2}–invariant. This means that there are local J_{S^2}–holomorphic coordinates centered at that point, in which $H_{S^2}(y) = (\text{constant}) \pm |y|^2 + O(|y|^3)$. When choosing inhomogeneous terms, we also require that they be zero at the local maxima and minima of H_{S^2}.

As a consequence of the condition on inhomogeneous terms, the constant map at a local minimum or maximum will be a solution of (7-8). One can use a counting-of-zeros argument for solutions of linear Cauchy–Riemann type operators on line bundles to show that the constant maps at minima have injective linearizations, and hence (since they have index 0) are regular. A similar argument, applied to (7-8) itself rather than its linearization, shows the following:

Lemma 7.6 (i) *Let $p \in S^2$ be a local maximum. For any solution of* (7-8) *on S^2, which is not constant equal to p, the homology class $[u] \in H_2(S^2) = \mathbb{Z}$ satisfies*

$$[u] \geq \#\{1 \leq i \leq d : x_i = p\} + \#\{u^{-1}(p)\}. \tag{7-25}$$

(ii) *Let $p \in S^2$ be a local minimum. For any solution of* (7-8) *on S^2, with $x_0 = p$, and which is not constant equal to p, we have*

$$[u] \geq 1 + \#\{u^{-1}(p)\}. \tag{7-26}$$

Take the graded abelian group obtained by using only critical points of H_{S^2} which have index ≤ 1 as generators. We want to equip this with a version of the quantum A_∞–structure, which uses inhomogeneous terms as in (7-24), but only considers solutions with degree $[u] = 0$. The argument showing that this works consists of three steps. First, Lemma 7.6(i) implies that for all solutions, one has $[u] \geq 0$. It follows that if we consider a sequence of maps of degree zero which converges to a limit with several pieces, then each piece must again have degree zero. Suppose that our original sequence consisted of maps whose limits are critical points of index ≤ 1. Our second point is that then, no critical point of index 2 can appear in the limit, since it would cause one of the pieces to have positive degree, again by Lemma 7.6(i). Thirdly, transversality of moduli spaces is unproblematic except possibly for the constant solutions at local minima; but we already know that such solutions are regular (in the ordinary sense of considering a fixed C, and therefore in the parametrized sense as well). We want to point out two properties of this A_∞–structure: the maps involved stay away from the local maxima, because of Lemma 7.6(i); and if u is a map that contributes to it, and whose limit x_0 is a local minimum, the map must actually be constant, by Lemma 7.6(ii) and the degree requirement.

Take a monotone symplectic manifold X. For each minimum or maximum p of H_{S^2}, we choose a Morse function and almost complex structure on X, written as $(H_{X,p}, J_{X,p})$. On the product $X \times S^2$, we then proceed as follows.

(7-27) Take a Hamiltonian $H_{X\times S^2}$ and almost complex structure $J_{X\times S^2}$, satisfying our usual conditions, and with the following additional properties. In a local J_{S^2}–holomorphic coordinate on S^2 around a local minimum or maximum p, we have $H_{X\times S^2} = H_{S^2} + H_{X,p} + O(|y|^3)$, and $J_{X\times S^2} = J_{S^2} \times J_{X,p} + O(|y|^2)$ similarly. When we choose inhomogeneous terms, they should have the property that, when restricted to $X \times \{p\}$, they take values in vector fields tangent to that submanifold.

As a consequence of this, we can have solutions of the associated equation (7-8) which are contained in $X \times \{p\}$. If p is a local minimum, then any such solution is regular in $X \times S^2$ if and only if it is regular inside $X \times \{p\}$. The counterpart of Lemma 7.6 for $X \times S^2$, proved by projecting to S^2 and arguing as before, is:

Lemma 7.7 (i) *Let $p \in S^2$ be a local maximum. For any solution of* (7-8) *on $X \times S^2$, which is not contained in $X \times \{p\}$, the homology class $[u] \in H_2(X \times S^2)$ satisfies*

$$[X]\cdot[u] \geq \#\{1 \leq i \leq d : x_i \in X \times \{p\}\} + \#\bigl\{u^{-1}(X \times \{p\})\bigr\}. \tag{7-28}$$

(ii) *Let $p \in S^2$ be a local minimum. For any solution of* (7-8) *on $X \times S^2$, with $x_0 \in X \times \{p\}$, and which is not contained in $X \times \{p\}$, we have*

$$[X]\cdot[u] \geq 1 + \#\bigl\{u^{-1}(X \times \{p\})\bigr\}. \tag{7-29}$$

For each local minimum p, we make choices of inhomogeneous terms which, building on the previously chosen $(H_{X,p}, J_{X,p})$, yield a quantum A_∞–structure $\mathcal{C}_p$. On $X \times S^2$, we then make corresponding choices, which restrict to the previous ones on $X \times \{p\}$ for each local minimum p. When building the corresponding version of the A_∞–structure on $X \times S^2$, denoted by $\mathcal{K}$, we use only those critical points of $H_{X \times S^2}$ which do not lie on $X \times \{p\}$ for a local maximum p; and only maps u with $[X]\cdot[u] = 0$. This works for exactly the same reasons as in the previously considered toy model case. Moreover, the following two properties hold: those maps that contribute avoid the subsets $X \times \{p\}$, where p is a local maximum; and projection to the subgroup generated by critical points in $X \times \{p\}$, where p is a local minimum, is a map

$$\mathcal{K} \to \mathcal{C}_p \tag{7-30}$$

compatible with the A_∞–structure. At this point, we specialize to functions H_{S^2} that have exactly one local maximum, but possibly several local minima. By looking at the Morse theory of $H_{X \times S^2}$, one sees that the projections (7-30) are chain homotopy equivalence. By looking at those maps for two local minima, one relates the A_∞–structures $\mathcal{C}_p$ for different p.

7e The formal group structure

Take the parameter spaces (3-9). We think of the interior of this space as parametrizing a family of punctured planes, which degenerate along the boundary. This is essentially constructed as in (3-13), but with two differences. First of all, we do include the spaces $\mathrm{MWW}_{0,\dots,1,\dots,0}$, to which we associate a once-punctured plane (a cylinder) with an inhomogeneous term, which is that defining the Floer differential. Since we are not dividing by translation, the only isolated point in the associated moduli space is a stationary solution at a critical point of the Hamiltonian, and that is the geometric origin of (4-28). Hence, to each "screen" in the limit corresponds a surface (unlike our original construction (3-13), where some of the screens were collapsed). The second difference is that we need everything to depend smoothly on parameters (the original construction was purely topological, hence allowed us to get away with continuity). More precisely, near the codimension-one boundary points of $\mathrm{MWW}_{d_1,\dots,d_r}$, we really need a situation as in Definition 7.3(ii); but along the codimension > 1 points, all we need is the situation from Definition 7.3(i), since those points only appear in compactness arguments. In any case, given the structure of $\mathrm{MWW}_{d_1,\dots,d_r}$ as a smooth manifold with generalized corners, it is unproblematic to define the required notion of smoothness, and to construct families of inhomogeneous terms satisfying it (by induction on dimension). The outcome are operations as in (4-26), with the difference that (4-28) is now a geometric statement rather than a separately imposed condition. One can therefore define (4-32) for the quantum A_∞–structure, and Lemmas 4.11–4.15 carry over immediately. What's important for applications is that we can, if desired, choose the inhomogeneous terms to be compatible with forgetting any color that has no marked points belonging to it; and therefore, to make our operations satisfy (4-54).

Well-definedness of (4-32) can be proved by a combination of the approaches from Sections 4d and 7c. Namely, suppose first that we have A_∞–rings $\mathcal{C}_0, \dots, \mathcal{C}_r$, each defined by a separate choice of function and other data. One can generalize (4-26) to obtain a map

$$\mathrm{MC}(\mathcal{C}_1; N) \times \cdots \times \mathrm{MC}(\mathcal{C}_r; N) \to \mathrm{MC}(\mathcal{C}_0; N). \tag{7-31}$$

Now, we want to change the A_∞–structure on $\mathcal{C}_0$ and one of the $\mathcal{C}_{k+1}$. The new versions are related to the old ones by larger A_∞–rings $\mathcal{H}_0$ and $\mathcal{H}_{k+1}$, constructed as in the uniqueness argument from Section 7c. The main tool in analyzing this change is an analogue of the middle $\to$ in (4-42), which is a map

$$\mathrm{MC}(\mathcal{C}_1; N) \times \cdots \times \mathrm{MC}(\mathcal{H}_{k+1}; N) \times \cdots \times \mathrm{MC}(\mathcal{C}_r; N) \to \mathrm{MC}(\mathcal{H}_0; N). \tag{7-32}$$

The definition of this involves two kinds of parameter spaces. The first ones are again the $\mathrm{MWW}_{d_1,\dots,d_r}$, but where we single out one of the d_{k+1} points of color $k+1$, as in the definition of (7-15), for special treatment when constructing the inhomogeneous terms and almost complex structures. The second class of parameter spaces are $\mathrm{MWW}_{d_1,\dots,d_{k+1}+1,\dots,d_r}$, which have an additional point of color $k+1$ (more precisely, there is one such space for every possible position of the additional point with respect to the other d_k). That point will not correspond to a puncture of the resulting Riemann surface; we just use its position as a modular variable, following the

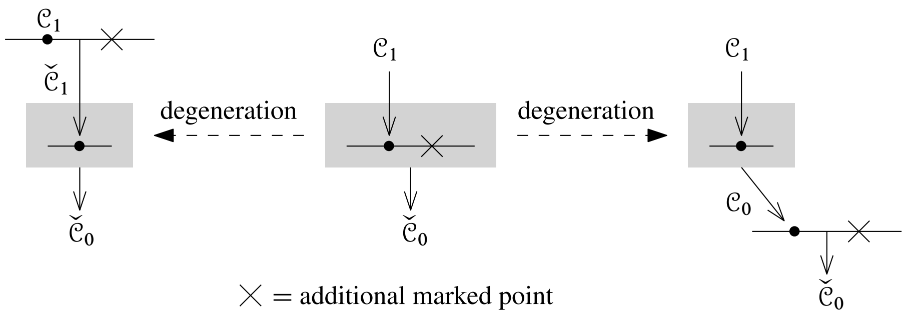

Figure 18: The construction from Example 7.8.

idea from (7-17). Of course, the additional marked point can in principle split off by itself into a mid-scale screen; when constructing the Riemann surface, that screen will not correspond to a component. We omit the details entirely.

Example 7.8 The simplest example of a parameter space of the second kind is MWW_{1+1}, where the additional marked point could be placed either on the left or right. The context in this case is that we have two versions of (7-21), namely $\mathcal{H}_k = \mathcal{C}_k u \oplus \widetilde{\mathcal{C}}_k \widetilde{u} \oplus \check{\mathcal{C}}_k v$ for $k = 0, 1$. As part of their A_∞–structure, we have continuation maps $\mathcal{C}_k \to \check{\mathcal{C}}_k$. On the other hand, as part of the $r = 1$ case of (7-31), we have constructed continuation maps $\mathcal{C}_1 \to \mathcal{C}_0$, $\widetilde{\mathcal{C}}_1 \to \widetilde{\mathcal{C}}_0$, $\check{\mathcal{C}}_1 \to \check{\mathcal{C}}_0$. The two versions of our moduli space then yield chain homotopies between compositions of those continuation maps, drawn as dashed arrows here:

$$\begin{tikzcd} \mathcal{C}_1 \arrow[r] \arrow[d] \arrow[dr, dashed] & \mathcal{C}_0 \arrow[d] \\ \check{\mathcal{C}}_1 \arrow[r] & \check{\mathcal{C}}_0 \end{tikzcd} \qquad \begin{tikzcd} \widetilde{\mathcal{C}}_1 \arrow[r] \arrow[d] \arrow[dr, dashed] & \widetilde{\mathcal{C}}_0 \arrow[d] \\ \check{\mathcal{C}}_1 \arrow[r] & \check{\mathcal{C}}_0 \end{tikzcd} \tag{7-33}$$

The geometry behind the construction on the left is shown schematically in Figure 18.

Remark 7.9 Alternatively, one could prove the well-definedness of the maps $\Pi^r_{\mathcal{C}}$ using the approach from Section 7d, which means defining a corresponding structure using $X \times S^2$, which comes with quasi-isomorphic projections to different copies of X.

The spaces SS from (3-27), while more complicated, show the same geometric behavior as MWW. Hence, the same kind of argument allows the proof of Proposition 4.20 to carry over, which completes our discussion of Proposition 1.5. There is a minor point which may be worth mentioning: in Section 3f, we added two marked points in the definition of the map (3-31), whose purpose was to break the symmetries of Fulton–MacPherson space. In a pseudoholomorphic curve context, we treat the extra points as in the well-definedness arguments above, meaning that their position gives additional modular variables on which the inhomogeneous term depends.

7f Commutativity

Adapting the arguments from Section 4f, we will now prove Proposition 1.6. This is the first time that one of the features of our Floer-theoretic setup, namely the time-independence of the Hamiltonians and almost complex structures, and the resulting S^1–symmetry of (7-1), will be used in a substantial manner.

Throughout the following discussion, it is assumed that choices of inhomogeneous terms have been made so as to satisfy (4-54). Let's start with the moduli space underlying $\beta_{\mathbb{C}}^{1,1}$. It involves a family of surfaces depending on one parameter, which we denote by $C_s = \mathbb{C} \setminus \{z_1(s), z_2(s)\}$ for $s \in \mathbb{R}$. One can assume that this family is symmetric outside a compact parameter range, in the sense that for some $S > 0$,

$$(z_1(-s), z_2(-s)) = (z_2(s), z_1(s)) \quad \text{if } |s| \geq S, \tag{7-34}$$

and that the inhomogeneous terms are chosen compatibly with this symmetry. As a consequence, there is partial cancellation between the two moduli spaces that enter into (4-61), with the parts having $|s| \geq S$ contributing only canceling pairs of points. One can therefore say that (4-61) is computed by a single parametrized moduli space, whose compact parameter space is a circle, obtained by gluing together the endpoints of two intervals $[-S, S]$. If we parametrize this circle by $r \in \mathbb{R}/2\pi\mathbb{Z}$ compatibly with its orientation, then that family of surfaces can be deformed to the simple form

$$(z_1(r), z_2(r)) = (\exp(r\sqrt{-1}), -\exp(r\sqrt{-1})). \tag{7-35}$$

Up to rotation, this is independent of r, and (it is here that we use time-independence) one can choose an inhomogeneous term to be compatible with that; in which case, the moduli space cannot have any isolated points, hence contributes zero. The deformation which ends up with (7-35), which can be thought as a family of surfaces parametrized by a compact two-dimensional disc, therefore gives rise to a nullhomotopy (4-65). As in our previous discussion of (4-60), this implies commutativity of the formal group structure mod N^3, which is the first part of Proposition 1.6.

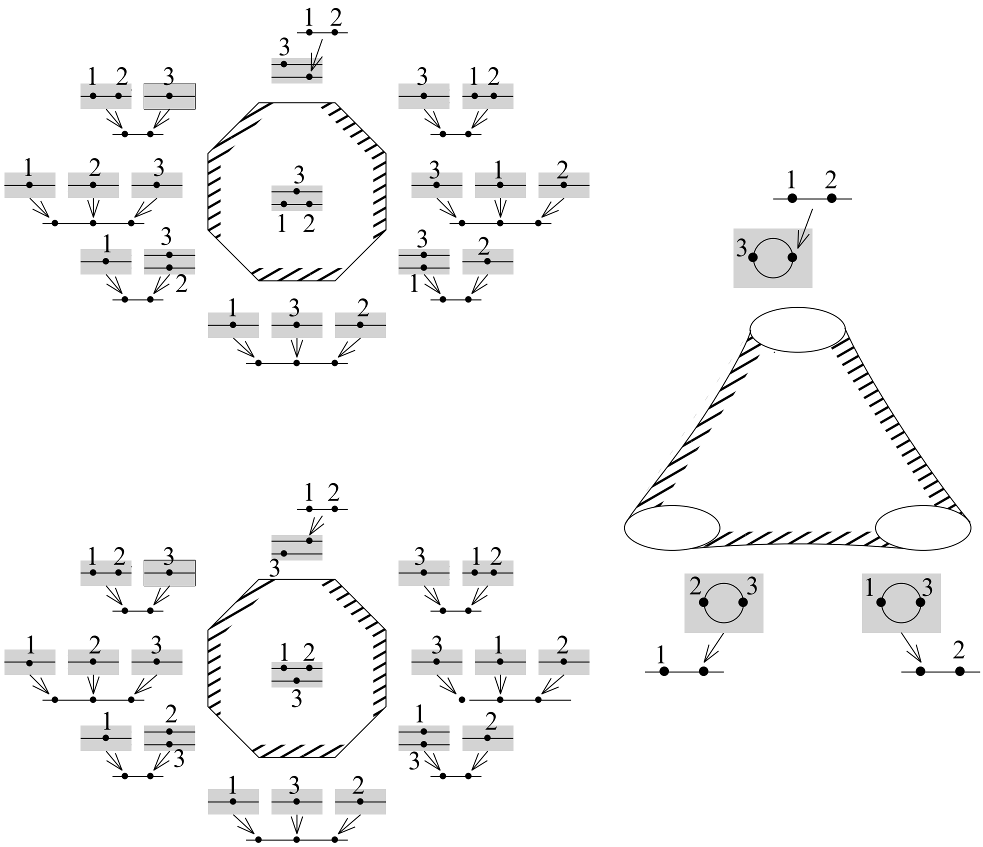

Figure 19: The geometry underlying (7-36): $\beta_{\mathcal{C}}^{2,1}(c_1, c_2; c_3)$, top left, and $\beta_{\mathcal{C}}^{1,2}(c_3; c_1; c_2)$, bottom left. Since part of the structure agrees, we can remove the hatched regions and join the rest together, with the outcome shown on the right after removing the "trivial screens" that have only one marked point. The pairs of points drawn as lying on a circle rotate around each other once in dependence on the parameters.

Next, let's look at part of the formula (4-69),

$$\beta_{\mathcal{C}}^{2,1}(c_1, c_2; c_3) - (-1)^{\|c_3\|(\|c_1\|+\|c_2\|)}\beta_{\mathcal{C}}^{1,2}(c_3; c_1, c_2). \tag{7-36}$$

The underlying moduli spaces are two copies of the octagon from Figure 5. Five of the boundary sides of those octagons match up in pairs which carry the same inhomogeneous term. We may assume that this extends to a neighborhood of those sides. As far as counting points in zero-dimensional moduli spaces is concerned, we can then cut out suitably matching neighborhoods and glue the rest together. The outcome of this process, shown in Figure 19, is that our expression can be computed by a single moduli space parametrized by a compact pair-of-pants surface. Moreover, along each boundary circle, we find that one of the components is a copy of the family of surfaces underlying (4-61), and which can be therefore filled in with a family parametrized by a disc. As a consequence, we find that the operation $K^{2,1}_{\mathcal{C}}$ defined in (4-69) is given by a family of surfaces parametrized by S^2. One can further deform that family so that degenerations happen only along three points (instead of the previous three discs) in the parameter space.

The outcome is that we have a family of four-punctured spheres, parametrized by S^2. Inspection of Figure 19 shows that this family has degree 1 in $H_2(\mathrm{DM}_3) \cong \mathbb{Z}$. This is a Floer-theoretic implementation of the four-pointed Gromov–Witten invariant, which we can relate to the standard version by a gluing argument as in [58]. As a consequence, identifying $H^*(\mathcal{C}) = H^*(X;\mathbb{Z})$, we have that on the cohomology level,

$$\int_X x_0\, K^{2,1}_{\mathcal{C}}(x_1, x_2; x_3) = \langle x_0, x_1, x_2, x_3\rangle_4, \quad \text{where } x_k \in H^*(X;\mathbb{Z}). \tag{7-37}$$

(See (9-1) for our notational conventions in Gromov–Witten theory.) For the particular case of $K^{2,1}_{\mathcal{C}}(x_1, x_1; x_2)$, where x_1 has odd degree, the graded symmetry of Gromov–Witten invariants means that $\langle x_0, x_1, x_1, x_2\rangle_4 = 0$. Assuming additionally that $H^*(X;\mathbb{Z})$ is torsion-free, it follows that $K^{2,1}_{\mathcal{C}}(x_1, x_1; x_2)$ itself is zero. As in Proposition 4.19, this and the corresponding argument for $K^{1,2}$ imply the desired commutativity statement modulo N^4.

7g The p^{th} power map

We now carry over the required arguments from Sections 5 and 6, leading to the proof of Theorem 1.9 as the Floer-theoretic analogue of Theorem 6.5.

We can bring Floer-theoretic constructions closer to the abstract operadic framework, by making a generic choice of inhomogeneous terms which are parametrized by FM_d. In the interior, this means that for every complex configuration $(z_1, \dots, z_d)$ we choose an inhomogeneous term ν_C on the resulting surface (7-4), in a way which is compatible with the action of the automorphisms which appear in (3-3). We then ask that this should extend to the "screens" associated to points in $\partial\mathrm{FM}_d$, in a way which enables compactness arguments for boundary strata of any dimension, see Definition 7.3(i), and gluing for codimension-one boundary strata, see Definition 7.3(ii). We also ask that our choices should be Sym_d–equivariant (recall that the symmetric group acts freely on FM_d).

Suppose that we have maps (3-13) based on smooth functions (3-15). By pull-back, our previous choice induces a family of inhomogeneous terms parametrized by $\mathrm{MWW}_{d_1,\dots,d_r}$. For a point in $\partial\mathrm{MWW}_{d_1,\dots,d_r}$, any vertices that are collapsed under (3-13) correspond to cylindrical components $C = \mathbb{C} \setminus \{z_1\}$, which we equip with the standard inhomogeneous term $\nu_C = \big(Z_H \otimes \mathrm{re}(d\log(z-z_1)/2\pi\sqrt{-1})\big)^{0,1}$. Moreover, a generic choice of (3-13) ensures transversality for the parametrized moduli spaces associated to all $\mathrm{MWW}_{d_1,\dots,d_r}$, and we may then use that choice to build the operations $\beta_{\mathcal{C}}^{d_1,\dots,d_r}$. Additionally, we may assume that the maps (3-13) are chosen so that (3-19) holds, which means that the resulting operations satisfy (4-54).

We will be specifically interested in $\mathrm{MWW}_{\underline{p}}$, in the notation from Section 6b, and the associated operation $\beta_{\mathcal{C}}^{\underline{p}} = \beta_{\mathcal{C}}^{1,\dots,1}$, with p prime. At this point, we fix an odd-degree cocycle

$$c \in \mathcal{C} \otimes \mathbb{F}_p, \tag{7-38}$$

which will remain the same throughout the subsequent discussion. Applying $\beta_{\mathcal{C}}^{\underline{p}}$ to p copies of c yields another such cocycle, hence a cohomology class

$$[\beta_{\mathcal{C}}^{\underline{p}}(c;\dots;c)] \in H^{\mathrm{odd}}(\mathcal{C};\mathbb{F}_p). \tag{7-39}$$

The underlying geometric phenomenon was explained in Section 6b: the codimension-one boundary faces of $\mathrm{MWW}_{\underline{p}}$ correspond to nontrivial decompositions of $\{1,\dots,p\}$ into nonempty subsets $(I_1,\dots,I_r)$ for any $r \geq 2$. If we act by an element of $\mathbb{Z}/p$ on such a decomposition, we get a new decomposition $(\tilde{I}_1,\dots,\tilde{I}_r)$, and the corresponding boundary faces, when mapped to FM_p, are related by the action of a suitable element of Sym_p; see (6-7). The cohomology class in (7-39) is independent of our choice of inhomogeneous terms.

Our first point is that we can realize (7-39) using a family of surfaces without degenerations. To do that, let's choose a Sym_p–equivariant isotopy that pushes Fulton–MacPherson space into its interior,

$$\begin{gathered} \phi_r : \mathrm{FM}_p \to \mathrm{FM}_p \quad \text{with } r \in [0,\epsilon], \\ \phi_0 = \mathrm{id}, \quad \phi_r(\mathrm{FM}_p) \subset \mathrm{FM}_p^\circ = \mathrm{FM}_p \setminus \partial\mathrm{FM}_p \quad \text{for } r > 0. \end{gathered} \tag{7-40}$$

Write $\iota_{\underline{p}}$ for the original map $\mathrm{MWW}_{\underline{p}} \to \mathrm{FM}_p$. The perturbed version,

$$\tilde{\iota}_{\underline{p}} = \phi_r \circ \iota_{\underline{p}} : \mathrm{MWW}_{\underline{p}} \to \mathrm{FM}_p \setminus \partial\mathrm{FM}_p \quad \text{for some } r > 0, \tag{7-41}$$

will retain the same $\mathbb{Z}/p$–action on codimension-one boundary faces as $\iota_{\underline{p}}$. Going back to the choice of inhomogeneous terms over FM_p, we want to also assume that the pullback of that family by (7-41) should lead to a regular parametrized moduli space. Given that, from (7-41) for some $r > 0$ we get a new operation $\tilde{\beta}^{p}_{\mathcal{C}}$, which again yields a cohomology class

$$[\tilde{\beta}^{p}_{\mathcal{C}}(c; \dots; c)] \in H^{\mathrm{odd}}(\mathcal{C}; \mathbb{F}_p). \tag{7-42}$$

A similar construction, where one interpolates between $\iota_{\underline{p}}$ and $\tilde{\iota}_{\underline{p}}$, shows that this cohomology class agrees with (7-39). At this point, we no longer need to compactify configuration space: to define (7-42), one can use families of perturbation data which are only defined on FM°_{p} (and still Sym_p–equivariant).

In the same vein as in (4-63), take

$$\mathcal{DM}^{\circ}_{p} = \mathrm{DM}^{\circ}_{p} \times E^{\circ}_{p}, \tag{7-43}$$

where E°_{p} is the interior of Fulton–MacPherson space for $\mathbb{R}^{\infty}$, meaning point configurations up to translation and rescaling. More precisely, we think of this as the direct limit of the corresponding spaces in each finite-dimensional Euclidean space. There is an embedding

$$\mathrm{FM}^{\circ}_{p} \to \mathcal{DM}^{\circ}_{p}, \tag{7-44}$$

which takes each point configuration to the pair formed by its quotient in Deligne–Mumford space and its image in E°_{p}. In this context, classes in $H^{\mathrm{Sym}_p}_{*}(\mathrm{DM}^{\circ}_{p}; \mathbb{F}_p(1))$ are realized by smooth simplicial chains in (7-43), having $\mathbb{F}_p$–coefficients, and quotiented out by the relation that acting on a chain by some $\sigma \in \mathrm{Sym}_p$ is the same as multiplying the chain with $(-1)^{\mathrm{sign}(\sigma)}$. Let's write $C^{\mathrm{Sym}_p}_{*}(\mathrm{DM}^{\circ}_{p}; \mathbb{F}_p(1))$ for this chain complex.

Choose a family of inhomogeneous terms on the family of surfaces pulled back by the projection $\mathcal{DM}^{\circ}_{p} \to \mathrm{DM}^{\circ}_{p}$, and which is Sym_p–equivariant. For every smooth map from a simplex to $\mathcal{DM}^{\circ}_{p}$, that family of inhomogeneous terms gives rise to a parametrized moduli space. If that space is regular, we get an operation $(\mathcal{C} \otimes \mathbb{F}_p)^{\otimes p} \to \mathcal{C} \otimes \mathbb{F}_p$. By adding up those operations with coefficients, we extend the construction to chains. Let's specialize to using p copies of our cocycle c as input. Morally, this can be thought of as giving rise to a $\mathbb{Z}/2$–graded chain map

$$C^{\mathrm{Sym}_p}_{*}(\mathrm{DM}^{\circ}_{p}; \mathbb{F}_p(1)) \to \mathcal{C}^{p|c|+*} \otimes \mathbb{F}_p. \tag{7-45}$$

The cautionary "morally" figures here because of the regularity condition for moduli spaces, which makes it impossible to define such a map on the entire chain complex. However, any argument involving a relation between specific chains, such as the one we are about to give, only involves finitely many terms, and one can assume that the chains involved are embedded into the infinite-dimensional space $\mathcal{DM}_p^\circ$. One can a posteriori make a choice of perturbation terms over $\mathcal{DM}_p^\circ$ which makes the finitely many spaces involved regular. Hence, for all practical purposes, the consequence is the same as if we had a map (7-45). In particular, we do get a map

$$H_*^{\mathrm{Sym}_p}(\mathrm{DM}_p^\circ; \mathbb{F}_p(1)) \to H^{p+*}(\mathcal{C}; \mathbb{F}_p). \tag{7-46}$$

One can think of (7-42) as an instance of this general construction, by smoothly triangulating the spaces $\mathrm{MWW}_{\underline{p}}$, in a way which is compatible with the $\mathbb{Z}/p$–action on codimension-one boundary strata, and then using the embedding (7-44). Using Lemma 5.3, one identifies the relevant homology class with that underlying the $t^{(p-1)/2}$ coefficient of the quantum Steenrod operation, up to a coefficient which is spelled out in Lemma 6.4. This equality, applied to (7-46), implies Theorem 1.9.

8 An alternative approach

The approach outlined in this section was pointed out to the author by Fukaya. It is an application of the results from [22] (taking the Lagrangian correspondence to be the diagonal, but with a general bounding cochain, which is our Maurer–Cartan element). The basic building blocks are parameter spaces from [47], which are close cousins of Stasheff associahedra (and in particular, are manifolds with corners in the classical sense). One can use them to define the composition law on Maurer–Cartan elements, a little indirectly, following [22, Theorem 1.7]; and to prove its associativity, following [22, Theorem 1.8]. On the other hand, it's not clear that there is a easier route from there to Theorem 1.9, which is one reason why we have not given first billing to this approach. Because of its complementary nature, our discussion will be quite sparse: not only are proofs omitted, we won't even make the distinction between the implementation of these arguments in an abstract operadic context (as in Section 4, assuming homological unitality) or a concrete Floer-theoretic one (as in Section 7).

8a The moduli spaces

We start with basically the same configuration space as in (3-10), except that the ordering of points in the last color is reversed:

$$\frac{\left\{(s_{1,1},\dots,s_{1,d_1};\dots;s_{r,1},\dots,s_{r,d_r}) : \begin{array}{l} s_{k,1}<\dots<s_{k,d_k} \text{ for } k<r, \\ s_{r,1}>\dots>s_{r,d_r} \end{array}\right\}}{\{s_{k,i}\sim s_{k,i}+\mu \text{ for } \mu\in\mathbb{R}\}}. \tag{8-1}$$

More importantly, we now consider a compactification of (8-1) which is smaller than its counterpart from Section 3d. This compactification will be denoted by

$$Q_{d_1,\dots,d_r}, \quad \text{where } r \geq 2,\ d_1,\dots,d_r \geq 0,\ d = d_1 + \dots + d_r > 0, \tag{8-2}$$

partly following the "quilted strips" terminology from [47]. The recursive structure of boundary strata is expressed by maps

$$\prod_{j=1}^{m} Q_{\|v_{\circ,j}\|_1,\dots,\|v_{\circ,j}\|_r} \times \prod_{v \text{ in } T_j,\ v \neq v_{\circ,j}} S_{\|v\|} \xrightarrow{(T_1,\dots,T_m)} Q_{d_1,\dots,d_r}. \tag{8-3}$$

Here, $T_1,\dots,T_m$ (for any $m \geq 1$) are trees of the following kind. In each T_j, denote by $v_{\circ,j}$ the vertex closest to the root. Then, the incoming edges at that vertex should carry one of r colors, and are ordered within their color. The parts of the tree lying above $v_{\circ,j}$ have planar embeddings, and inherit a single color. The whole thing is arranged, of course, so that the total number of leaves of each respective color add up to $(d_1,\dots,d_r)$. Geometrically, what happens is that as groups of points move to $\pm\infty$, we split them up into separate screens, which correspond to the Q factors in (8-3) (in the terminology of Section 3d, these would be called mid-scale, since there is no rescaling involved, just translation); but we do not keep track of the relative speeds at which this divergence happens (no large-scale screens). The remaining factors in (8-3) are small-scale screens, which describe the limit of points converging towards each other. The image of (8-3) has codimension equal to the overall number of factors (vertices) minus one. This is related to the fact that $Q_{d_1,\dots,d_r}$ is a smooth manifold with corners.

Let's map our points to radial half-lines in the punctured plane,

$$z_{k,i} = \exp\Bigl(-s_{k,i} - \frac{2\pi k}{r}\sqrt{-1}\Bigr) \in \mathbb{C}^*, \tag{8-4}$$

and add a marked point at 0 (the $z_{k,i}$ are ordered lexicographically, and the extra point is inserted between the last two colors). This extends to a continuous map

$$Q_{d_1,\dots,d_r} \to \mathrm{FM}_{d+1}. \tag{8-5}$$

In terms more familiar from pseudoholomorphic curve theory, one can think of the configurations (8-4) as lying on parallel lines on a cylinder. In the limit, this breaks up into several cylinders, plus spheres (copies of $\overline{\mathbb{C}}$ with a marked point at infinity, and other marked points lying on the real line) attached to them; see Figures 20(i) and 21(i). On the combinatorial level, the map (8-5) works as follows: starting with trees as in (8-3), one adds an incoming edge to each vertex $v_{\circ,j}$ except the last one, and then identifies those edges with the root edges of T_{j+1}, thereby combining all our trees into a single T, which is what appears in (3-4); see Figure 20(ii).

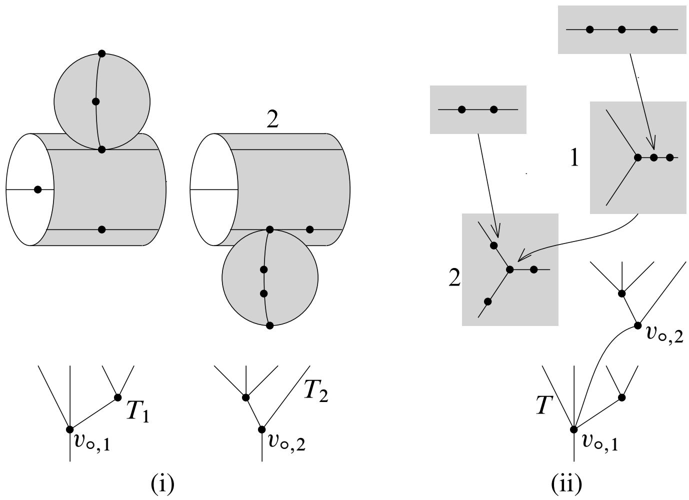

Figure 20: (i) A boundary point in (8-2), drawn in the way familiar from pseudoholomorphic curve theory; and (ii) the corresponding picture in Fulton–MacPherson space.

8b The operations

Algebraically, the outcome of using (8-2) and (8-5) are operations

$$\chi_{\mathcal{C}}^{d_1,\dots,d_{r-1},1,d_r} : \mathcal{C}^{d_1+\dots+d_r+1} \to \mathcal{C}[1-d_1-\dots-d_r]. \tag{8-6}$$

Additionally, we set

$$\chi_{\mathcal{C}}^{0,\dots,1,0} = \mu_{\mathcal{C}}^1. \tag{8-7}$$

The property of these operations, derived as usual from the structure of codimension-one boundary strata, and including (8-7), is that

$$\begin{aligned}
\sum_{\substack{k<r\\ ij}} \pm \chi_{\mathcal{C}}^{d_1,\dots,d_k-j+1,\dots,1,d_r}\big(&c_{1,1},\dots,c_{1,d_1};\dots;c_{k,1},\dots,\\
&\mu_{\mathcal{C}}^j(c_{k,i+1},\dots,c_{k,i+j}),\dots,c_{k,d_k};\dots;c;c_{r,1},\dots,c_{r,d_r}\big)\\
+\sum_{p_1,\dots,p_r} \pm \chi_{\mathcal{C}}^{d_1-p_1,\dots,1,d_r-p_r}\big(&c_{1,1},\dots,c_{1,p_1};\dots;c_{r-1,1},\dots,c_{r,p_{r-1}};\\
&\chi_{\mathcal{C}}^{p_1,\dots,1,p_r}(c_{1,p_1+1},\dots,c_{1,d_1};\dots;c;c_{r,p_1},\dots,c_{r,p_r});\\
&c_{r,p_r+1},\dots,c_{r,d_r}\big)\\
+\sum_{ij} \pm \chi_{\mathcal{C}}^{d_1,\dots,1,d_r-j+1}\big(&c_{1,1},\dots;c;c_{r,1},\dots,c_{r,i},\mu_{\mathcal{C}}^j(c_{r,i+1},\dots,c_{r,i+j}),\\
&\dots,c_{r,d_r}\big)\\
&=0.
\end{aligned} \tag{8-8}$$

In terminology similar to [47], these define the structure of an A_∞–$(r-1,1)$–module, with the first $r-1$ factors acting on the left, and the last one on the right.

Example 8.1 Suppose that $d_1 = \cdots = d_{r-1} = 0$. Then, the points (8-4), with the origin added as usual, lie on a half-line in $\mathbb{C}$. One can use that to identify $Q_{0,\dots,0,d} \cong S_{d+1}$. The auxiliary data involved in defining (8-6) can be chosen to be compatible with that, in which case one gets $\chi_{\mathcal{C}}^{0,\dots,1,d} = \mu_{\mathcal{C}}^{d+1}$.

Example 8.2 For $r=2$, the points (8-4) still lie on $\mathbb{R} \subset \mathbb{C}$, hence $Q_{d_1,d_2} \cong S_{d_1+d_2+1}$ and, for suitable choices, $\chi_{\mathcal{C}}^{d_1,1,d_2} = \mu_{\mathcal{C}}^{d_1+d_2+1}$.

Example 8.3 The operations with two inputs, $\chi_{\mathcal{C}}^{0,\dots,1,\dots,1,0}$ and $\chi_{\mathcal{C}}^{0,\dots,1,1}$, are all chain homotopic to the multiplication $\mu_{\mathcal{C}}^2$, simply because they come from a single two-point configuration in the plane.

Our purpose in defining these operations is the following:

Definition 8.4 Let $\gamma_1,\dots,\gamma_r \in \mathcal{C}^1 \mathbin{\hat{\otimes}} N$ be Maurer–Cartan elements. We say that γ_r is the product of $(\gamma_1,\dots,\gamma_{r-1})$ if there is a $k \in \mathcal{C}^0 \mathbin{\hat{\otimes}} (\mathbb{Z}1 \oplus N)$ which modulo N reduces to a cocycle representing the unit $[e_{\mathcal{C}}]$, and such that

$$\text{(8-9)}\qquad \sum_{d_1,\dots,d_r} \chi_{\mathcal{C}}^{d_1,\dots,d_{r-1},1,d_r}(\overbrace{\gamma_1,\dots,\gamma_1}^{d_1};\dots;\overbrace{\gamma_{r-1},\dots,\gamma_{r-1}}^{d_{r-1}};k;\overbrace{\gamma_r,\dots,\gamma_r}^{d_r}) = 0.$$

The expression (8-9) includes a term $\mu_{\mathcal{C}}^1(k)$, corresponding to $(d_1,\dots,d_r) = (0,\dots,0)$. If write $k = e_{\mathcal{C}} + (\text{coboundary}) + h$ with $h \in \mathcal{C}^0 \mathbin{\hat{\otimes}} N$, then (keeping Example 8.3 in mind) the next-order term in the equation says that

$$\begin{aligned} \text{(8-10)}\quad & [\chi_{\mathcal{C}}^{1,0,\dots,1,0}(\gamma_1,k)] + \cdots + [\chi_{\mathcal{C}}^{0,\dots,1,1,0}(\gamma_{r-1},k)] + [\chi_{\mathcal{C}}^{0,\dots,1,1}(k,\gamma_r)] \\ &= [\mu_{\mathcal{C}}^2(\gamma_1,e_{\mathcal{C}})] + \cdots + [\mu_{\mathcal{C}}^2(\gamma_{r-1},e_{\mathcal{C}})] + [\mu_{\mathcal{C}}^2(e_{\mathcal{C}},\gamma_r)] \\ &= [-\gamma_1 - \cdots - \gamma_{r-1} + \gamma_r] = 0 \qquad \text{in } H^1(\mathcal{C} \otimes N/N^2). \end{aligned}$$

For $r=2$, and assuming the choices have been made as in Example 8.2, the condition in (8-9) reduces to the criterion for equivalence of γ_1 and γ_2 given in Lemma 2.6.

Lemma 8.5 *The notion of product from Definition 8.4 only depends on the equivalence class of the Maurer–Cartan elements involved.*

This is the analogue of Lemma 4.11, and is proved in a similar way. Given $(\gamma_1,\dots,\gamma_r)$ and k as in (8-9), and an element $h \in \mathcal{C}^0 \mathbin{\widehat{\otimes}} N$ which provides an equivalence between γ_j and $\tilde{\gamma}_j$, we can construct an explicit $\tilde{k}$ which shows that $(\gamma_1,\dots,\tilde{\gamma}_j,\dots,\gamma_r)$ satisfy the same condition:

$$\tilde{k} = k + \sum_{d_1,\dots,d_r,i} \chi_{\mathcal{C}}^{d_1,\dots,1,d_r}(\dots;\overbrace{\gamma_j,\dots,\gamma_j}^{i},h,\overbrace{\tilde{\gamma}_j,\dots,\tilde{\gamma}_j}^{d_j-i-1};\dots;k;\dots). \tag{8-11}$$

(The formula as written is for $j<r$, but the $j=r$ case is parallel.)

Lemma 8.6 *Given $(\gamma_1,\dots,\gamma_{r-1})$, there is a unique equivalence class γ_r satisfying Definition 8.4.*

This is roughly analogous to Lemma 4.14. It is maybe helpful to reformulate the issue as follows. We have a right A_∞–module structure, defined by

$$\begin{aligned}(c;c_1,\dots,c_d) \\ \mapsto \sum_{d_1,\dots,d_{r-1}} \chi_{\mathcal{C}}^{d_1,\dots,d_{r-1},1,d}(\overbrace{\gamma_1,\dots,\gamma_1}^{d_1};\dots;\overbrace{\gamma_{r-1},\dots,\gamma_{r-1}}^{d_{r-1}};c;c_1,\dots,c_d),\end{aligned} \tag{8-12}$$

which (thanks to Example 8.1) is a deformation of the free module $\mathcal{C}$. One then wants to modify that module structure through γ insertions, so as to "undo" the deformation, rendering it trivial. This is a purely algebraic question, which can be reduced to the strictly unital situation if desired (using Lemma 2.2).

Proposition 8.7 *In the sense of Definition 8.4, if γ is the product of (γ_i,γ_{i+1}) for some $i<r-1$, and γ_r is the product of $(\gamma_1,\dots,\gamma_{i-1},\gamma,\gamma_{i+2},\dots,\gamma_{r-1})$, then γ_r is also the product of $(\gamma_1,\dots,\gamma_{r-1})$.*

This is the associativity statement for our notion of product. The proof uses a moduli space of points lying on certain lines in the punctured plane. It is convenient to draw that plane as a pair-of-pants, see Figure 21(ii), which is half of the "double pants diagram" in [22, Section 11.2]. The two known statements about products come with their respective elements k as in (8-9). One inserts those elements at the two bottom ends, and the Maurer–Cartan elements at points on the respective lines (the arrows denote the ordering of the points), the outcome being another element k which establishes the desired statement.

Remark 8.8 Let's briefly discuss the counterpart of Definition 8.4 in homological algebra, in the spirit of Section 2. Take a homologically unital A_∞–ring $\mathcal{A}$, and its

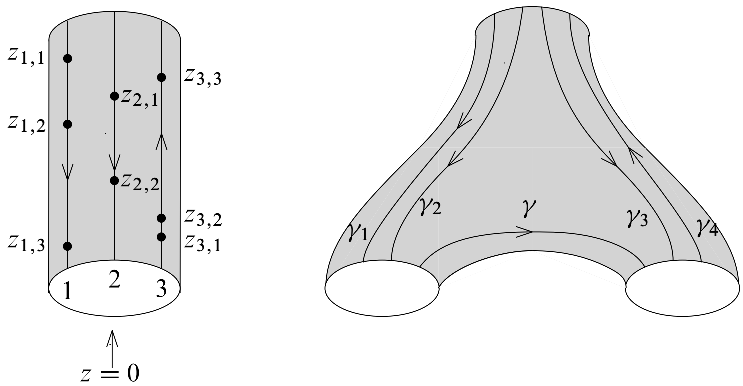

Figure 21: (i) Another picture of a point configuration (8-4), with correct ordering of the lines and points. (ii) An analogous picture of the moduli spaces that enter into the proof of Proposition 8.7 ($r = 4$, $i = 1$).

Hochschild complex $\mathcal{C}$; see (2-13). Given a Maurer–Cartan element $\gamma \in \mathcal{C}^1 \mathbin{\hat{\otimes}} N$, one can define two A_∞–bimodules $\mathcal{L}_\gamma$ and $\mathcal{R}_\gamma$, whose underlying space is $\mathcal{A} \mathbin{\hat{\otimes}} (\mathbb{Z}1 \oplus N)$, with bimodule structure

$$
\begin{aligned}
&\mu^{p;1;q}_{\mathcal{L}_\gamma}(a_1,\dots,a_p;a_{p+1};a_{p+2},\dots,a_{p+q+1}) = \\
&\pm\mu^{p+q+1}_{\mathcal{A}}(a_1,\dots,a_{p+q+1}) \\
&\quad + \sum_{ij} \pm\, \mu^{p+q+2-j}_{\mathcal{A}}\bigl(a_1,\dots,\dots,a_i,\gamma^j(a_{i+1},\dots,a_{i+j}),\dots,a_{p+1}, \\
&\hspace{20em} \dots,a_{p+q+1}\bigr), \\
&\quad + \sum_{i_1 j_1 i_2 j_2} \pm\, \mu^{p+q+3-j_1-j_2}_{\mathcal{A}}\bigl(a_1,\dots,a_{i_1},\gamma^{j_1}(a_{i_1+1},\dots,a_{i_1+j_1}), \\
&\hspace{8em} \dots,a_{i_2},\gamma^{j_2}(a_{i_2+1},\dots,a_{i_2+j_2}),\dots,a_{p+1},\dots,a_{p+q+1}\bigr) \\
&\hspace{24em} + \cdots
\end{aligned} \tag{8-13}
$$

together with

$$
\begin{aligned}
&\mu^{p;1;q}_{\mathcal{R}_\gamma}(a_1,\dots,a_p;a_{p+1};a_{p+2},\dots,a_{p+q+1}) = \\
&\pm\mu^{p+q+1}_{\mathcal{A}}(a_1,\dots,a_{p+q+1}) \\
&+\sum_{ij} \pm\mu^{p+q+2-j}_{\mathcal{A}}(a_1,\dots,a_{p+1},\dots,a_i,\gamma^j(a_{i+1},\dots,a_{i+j}),\dots,a_{p+q+1}) \\
&+ \sum_{i_1 j_1 i_2 j_2} \pm\, \mu^{p+q+3-j_1-j_2}_{\mathcal{A}}\bigl(a_1,\dots,a_{p+1},\dots,a_{i_1},\gamma^{j_1}(a_{i_1+1},\dots,a_{i_1+j_1}), \\
&\hspace{12em} \dots,a_{i_2},\gamma^{j_2}(a_{i_2+1},\dots,a_{i_2+j_2}),\dots,a_{p+q+1}\bigr) \\
&\hspace{24em} + \cdots .
\end{aligned} \tag{8-14}
$$

Here, the rule is that an arbitrary number of γ terms are inserted, but always to the left (8-13), or right (8-14), of a_{p+1}. For $\gamma = 0$, this reduces to $\mathcal{A}$ with the diagonal bimodule structure extended to $\mathcal{A}\hat{\otimes}(\mathbb{Z}1 \oplus N)$, which will denote by $\mathcal{L}_0 = \mathcal{R}_0 = \mathcal{D}$. More generally, $\mathcal{L}_\gamma$ and $\mathcal{R}_\gamma$ can be viewed as pullbacks of $\mathcal{D}$ by the formal automorphism (2-23) acting on one of the two sides. Using that, one sees easily that the bimodules are inverses: there are bimodule homotopy equivalences

$$\mathcal{R}_\gamma \otimes_{\mathcal{A}} \mathcal{L}_\gamma \simeq \mathcal{L}_\gamma \otimes_{\mathcal{A}} \mathcal{R}_\gamma \simeq \mathcal{D}. \tag{8-15}$$

Here, the tensor product notation is shorthand: we are really taking the tensor product of A_∞–bimodules relative to $\mathcal{A} \mathbin{\hat{\otimes}} (\mathbb{Z}1 \oplus N)$, and making sure that completion with respect to the filtration of N is taken into account. Modulo N, all our bimodules reduce to the diagonal bimodule. Consider the Hochschild complex of $\mathcal{A}$ with coefficients in a bimodule $\mathcal{B}$, denoted here by $\mathrm{CC}^*(\mathcal{B})$; see eg [26, Section 2.9]. We say that γ_r is the product of $(\gamma_1, \dots, \gamma_{r-1})$ if there is a cocycle

$$k \in \mathrm{CC}^0(\mathcal{R}_{\gamma_1} \otimes_{\mathcal{A}} \cdots \otimes_{\mathcal{A}} \mathcal{R}_{\gamma_{r-1}} \otimes_{\mathcal{A}} \mathcal{L}_{\gamma_r}) \tag{8-16}$$

which, after reduction modulo N, represents the identity in $HH^0(\mathcal{A})$. For $r = 1$, one can use (8-15) to show that this is the case if and only if γ_1, γ_2 are equivalent. On the other hand, for $\bullet$ defined as in (2-24), there are homotopy equivalences

$$\mathcal{L}_{\gamma_1} \otimes_{\mathcal{A}} \cdots \otimes_{\mathcal{A}} \mathcal{L}_{\gamma_{r-1}} \simeq \mathcal{L}_{\gamma_{r-1} \bullet \cdots \bullet \gamma_1} \quad \text{and} \quad \mathcal{R}_{\gamma_1} \otimes_{\mathcal{A}} \cdots \otimes_{\mathcal{A}} \mathcal{R}_{\gamma_{r-1}} \simeq \mathcal{R}_{\gamma_1 \bullet \cdots \bullet \gamma_{r-1}}. \tag{8-17}$$

Combining that with the previous observation shows that our definition of product is the same as saying that γ_r is equivalent to $\gamma_1 \bullet \cdots \bullet \gamma_{r-1}$.

8c Relating the two approaches

To conclude our discussion, we'll mention a possible way to connect the construction in this section to the rest of the paper, or more precisely: to prove that, if γ_r is the product of $(\gamma_1, \dots, \gamma_{r-1})$ in the sense of Definition 8.4, then γ_r is also equivalent to $\Pi_{\mathcal{C}}^{r-1}(\gamma_1, \dots, \gamma_{r-1})$, which is the product from Definition 4.10. The reader is exhorted to treat this as what it is, a suggestion: it relies on new moduli spaces whose structure has not been fully developed.

We begin by introducing an additional parameter $t \in [0, 1)$, and changing (8-4) by letting the first $r-1$ radial lines collide in the limit $t \to 1$:

$$z_{k,i} = \begin{cases} \exp\Bigl(-s_{k,i} - (1-t)\dfrac{2\pi k}{r}\sqrt{-1} - t\pi\sqrt{-1}\Bigr) \in \mathbb{C}^* & \text{if } k = 1, \dots, r-1, \\ \exp(-s_{r,i}) & \text{if } k = r. \end{cases} \tag{8-18}$$

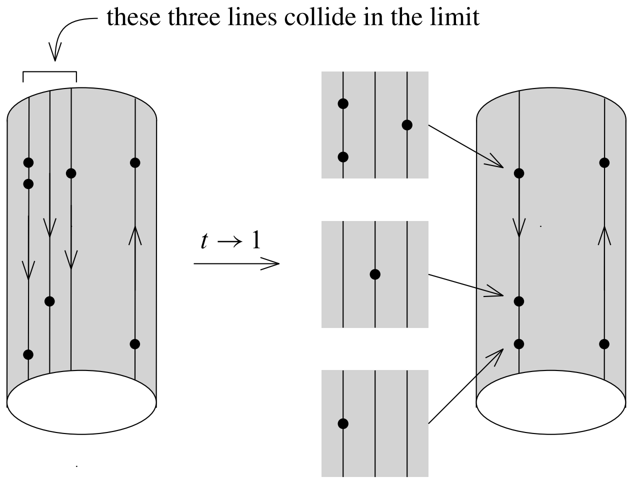

Figure 22: A limit in the space $PQ_{3,1,1,2}$ which lies in the image of (8-20). For compatibility with Figure 21, we have drawn the MWW components rotated by 90 degrees.

This leads to a compactification of $[0, 1)$ times (8-1), which we write as

$$t\colon PQ_{d_1,\dots,d_r} \to [0,1] \quad \text{for } r \geq 3. \tag{8-19}$$

Over each $t \in [0, 1)$, the fiber is a copy of $Q_{d_1,\dots,d_r}$. In the limit $t \to 1$, points of the first $r-1$ colors bubble off into screens which have the structure of MWW spaces (Figure 22). This means that we have maps

$$Q_{j,r_d} \times \prod_{i=1}^{j} \mathrm{MWW}_{d_{1,i},\dots,d_{r-1,i}} \to PQ_{d_1,\dots,d_r} \tag{8-20}$$

for each partition $d_1 = d_{1,1} + \cdots + d_{j,1}, \dots, d_{r-1} = d_{r-1,1} + \cdots + d_{r-1,j}$, whose images are the top-dimensional parts of the fiber of (8-19) over $t = 1$. The space (8-19) comes with a map to FM_{d+1}, which over the fiber $t = 0$ reduces to (8-5).

The definition (8-18) suffers from the usual disadvantage of parametrized spaces, meaning that its compactification contains strata that are fiber products over $[0, 1]$. To bypass that difficulty, one can try to use those spaces in a "time-ordered" form (the same strategy as in [65, Section 10e]), which means that we consider k–tuples of points in $PQ_{d_{1,1},\dots,d_{1,r}} \times \cdots \times PQ_{d_{k,1},\dots,d_{k,r}}$, where $(d_{i,k})$ is again a partition of (d_k), such that the associated parameters satisfy $t_1 \leq \cdots \leq t_k$. Pairs of points with equal parameters now occur in the boundary of two different such spaces (when $t_i = t_{i+1}$

for some i; and as a boundary face of one of the PQ factors involved). If we insert Maurer–Cartan elements $(\gamma_1, \dots, \gamma_r)$ at the marked points in $\mathbb{C}^*$, and add a trivial term which is the identity map, the outcome should be a map

$$\Phi_{\mathcal{C}}\colon \mathcal{C} \to \mathcal{C} \tag{8-21}$$

which, setting $\gamma = \Pi_{\mathcal{C}}^{r-1}(\gamma_1, \dots, \gamma_{r-1})$, satisfies

$$\Phi_{\mathcal{C}}\Bigl(\sum_{d_1,\dots,d_r} \chi_{\mathcal{C}}^{d_1,\dots,d_{r-1},1,d_r}(\overbrace{\gamma_1, \dots, \gamma_1}^{d_1}; \dots; \overbrace{\gamma_{r-1}, \dots, \gamma_{r-1}}^{d_{r-1}}; c; \overbrace{\gamma_r, \dots, \gamma_r}^{d_r})\Bigr) \\ = \sum_{d_1,d_2} \chi_{\mathcal{C}}^{d_1,1,d_2}(\underbrace{\gamma, \dots, \gamma}_{d_1}; \Phi_{\mathcal{C}}(c); \underbrace{\gamma_r, \dots, \gamma_r}_{d_2}). \tag{8-22}$$

The left-hand side of this equation represents what happens for $t_1 = 0$ (the $t_1 = 0$ component then gives $\chi_{\mathcal{C}}$, and the other components give $\Phi_{\mathcal{C}}$), while the right-hand side represents what happens for $t_k = 1$ (with the γ factors coming from the collision of the first $r-1$ lines). Suppose that γ_r is the product of $(\gamma_1, \dots, \gamma_{r-1})$; see Definition 8.4. Then, inserting the associated element k into (8-21) produces another element, which shows that "γ_r is the product of (γ)" in the same sense. As pointed out before, in that special case, the definition just amounts to saying that γ_r and γ are equivalent.

9 Computing quantum Steenrod operations

By definition, quantum Steenrod operations belong to genus-zero enumerative geometry. Generally speaking, it's an open question what their role is within that theory. However, for low-degree contributions one can give a satisfactory answer, in terms of the usual Gromov–Witten invariants. After explaining this, we will turn to specific example computations.

9a Gromov–Witten theory background

Let's start in a context which is a little different than the rest of the paper. Take X to be a closed symplectic $2n$–manifold, with the only restriction (for notational simplicity, since we only want to use power series in the Novikov variable q) that the symplectic form must lie in an integral cohomology class, denoted here by

$$\Omega_X \in \operatorname{Im}\bigl(H^2(X;\mathbb{Z}) \to H^2(X;\mathbb{Q})\bigr).$$

Genus-zero Gromov–Witten invariants for m–pointed curves, and their generalizations that include gravitational descendants, will be written as

$$\begin{aligned}\text{(9-1)}\quad &\langle \psi^{r_1}x_1,\dots,\psi^{r_m}x_m\rangle_m \\ &\qquad = \sum_A q^{\Omega_X\cdot A}\langle \psi^{r_1}x_1,\dots,\psi^{r_m}x_m\rangle_{m,A} \in \mathbb{Q}[\![q]\!] \quad \text{for } x_i\in H^*(X;\mathbb{Q}),\end{aligned}$$

where the sum is over $A\in H_2(X;\mathbb{Z})$. For the contribution of A to be potentially nonzero, one should either consider classes with positive symplectic area $\Omega_X\cdot A=\int_A \omega_X>0$, or take $A=0$ (the case of constant curves) and $m\ge 3$. For expositions of Gromov–Witten that include the properties used in this paper, see eg [57, Section 1] or [35, Chapter 26].

We introduce another formal variable t, so that the coefficient ring for our algebraic considerations will be $\mathbb{Q}[t^{\pm 1}][\![q]\!]$. The small quantum product, and the small quantum connection, on the $\mathbb{Z}/2$–graded space $H^*(X;\mathbb{Q})[t^{\pm1}][\![q]\!]$ are defined by

$$\int_X (y_1 * y_2)\, y_3 = \langle y_1,y_2,y_3\rangle_3, \tag{9-2}$$

$$\nabla y = q\partial_q y + t^{-1}\Omega_X * y. \tag{9-3}$$

We will consider endomorphisms Φ of $H^*(X;\mathbb{Q})[t^{\pm1}][\![q]\!]$ which are (linear over the coefficient ring and) covariantly constant with respect to ∇. Concretely, this means that

$$(q\partial_q\Phi)(y) + t^{-1}\Omega_X * \Phi(y) - t^{-1}\Phi(\Omega_X * y) = 0. \tag{9-4}$$

If we expand $\Phi = \Phi^{(0)} + q\Phi^{(1)} + q^2\Phi^{(2)} + \cdots$, (9-4) becomes

$$\Phi^{(0)}(\Omega_X y) = \Omega_X \Phi^{(0)}(y), \tag{9-5}$$

$$\Phi^{(k)}(y) = t^{-1}k^{-1}(\Phi^{(k)}(\Omega_X y) - \Omega_X\Phi^{(k)}(y)) + (\text{recursive terms}) \quad \text{if } k>0, \tag{9-6}$$

where the generic "recursive terms" covers expressions involving only $\Phi^{(0)},\dots,\Phi^{(k-1)}$. By repeatedly inserting (9-6) into itself, we get

$$\Phi^{(k)}(y) = t^{-m}k^{-m}\sum_i (-1)^i\binom{m}{i}\Omega_X^i\Phi^{(k)}(\Omega_X^{m-i}y) + (\text{recursive terms}). \tag{9-7}$$

Setting $m>2n$ means that in the sum we have $i>n$ or $m-i>n$, so all those terms vanish. One therefore gets explicit recursive formulae, which show that the constant term $\Phi^{(0)}$, subject to (9-5), determines all of Φ. The case we are interested in is where $\Phi^{(0)}(y)=xy$ is the cup product with a given class $x\in H^*(X;\mathbb{Q})$. There is a formula for the resulting $\Phi=\Phi_x$ in terms of gravitational descendants, closely related to the standard formula for solutions of the quantum differential equation:

$$
\begin{aligned}
\text{(9-8)}\quad & \int_X y_0\,\Phi_x(y_1) \\
&= \int_X y_0 x y_1 - t^{-1}\langle y_0, (1+t^{-1}\psi)^{-1} x y_1\rangle_2 + t^{-1}\langle (1-t^{-1}\psi)^{-1} y_0 x, y_1\rangle_2 \\
&\qquad - t^{-2}\sum_k \langle y_0, (1+t^{-1}\psi)^{-1} x e_k\rangle_2 \langle (1-t^{-1}\psi)^{-1} e_k^\vee, y_1\rangle_2.
\end{aligned}
$$

In principle, the terms $(1+\cdots)^{-1}$ are supposed to be expanded into geometric series; but for degree reasons, only one term in this series is nonzero for each class A that contributes to the expressions in (9-8). The (e_k), $(e_k^\vee)$ are Poincaré dual bases in $H^*(X;\mathbb{Q})$, meaning that in the Künneth decomposition,

$$
\text{(9-9)}\qquad \sum_k e_k \otimes e_k^\vee = [\text{diagonal}] \in H^{2n}(X\times X;\mathbb{Q}).
$$

Checking that Φ_x satisfies (9-4) is an exercise using basic properties (divisor equation and TRR) of Gromov–Witten invariants. Using the string equation, one can write the special case $y_0 = y$, $y_1 = 1$ as

$$
\begin{aligned}
\text{(9-10)}\quad \int_X y\,\Phi_x(1) = \int_X yx - t^{-1}\langle y, (1+t^{-1}\psi)^{-1} x\rangle_2 + t^{-2}\langle (1-t^{-1}\psi)^{-1} yx\rangle_1 \\
- t^{-3}\sum_k \langle y, (1+t^{-1}\psi)^{-1} x e_k\rangle_2 \langle (1-t^{-1}\psi)^{-1} e_k^\vee\rangle_1.
\end{aligned}
$$

Let's modify the context slightly, and assume that X is weakly monotone. Moreover, choose an integer lift of the symplectic cohomology class, again denoted by Ω_X. Then, one can define mod p versions of Gromov–Witten invariants counting curves in $A \in H_2(X;\mathbb{Z})$, for which we use the same notation:

$$
\text{(9-11)}\qquad \langle x_1,\dots,x_m\rangle_{m,A} \in \mathbb{F}_p \quad \text{with } x_i \in H^*(X;\mathbb{F}_p), \qquad \text{provided that } \begin{cases} m \geq 3, \text{ or} \\ \text{any } m \text{ and } 0 < \Omega_X\cdot A < p. \end{cases}
$$

For $m \geq 3$, this is the classical definition in terms of an inhomogeneous $\bar\partial$–equation (Gromov's trick). The definition in the second case can be reduced to the first case by taking the divisor equation as an axiom, where the class inserted is always (the mod p reduction of) Ω_X. Alternatively, one could argue more geometrically: if $\Omega_X\cdot A < p$, then no stable map in class A can have an automorphism group whose order is a multiple of p. This should allow one to define virtual fundamental classes in homology with $\mathbb{F}_p$–coefficients (we say "should" since this has not, to our knowledge, been carried out in the literature). The discussion of the second case also applies to gravitational descendants, with the same assumption $0 < \Omega_X \cdot A < p$. Geometrically, this uses the fact that orbifold line bundles whose isotropy groups have orders coprime to p have Chern classes in mod p cohomology; algebraically, one can use the formula (involving the divisor relation and TRR) that reduces invariants involving gravitational descendants to ordinary Gromov–Witten invariants.

The quantum product and connection can be considered as acting on $H^*(X;\mathbb{F}_p)[t^{\pm 1}][\![q]\!]$, where one now thinks of t as in (1-5). Formal linear differential equations in characteristic p have a much larger space of solutions than their characteristic 0 counterparts, simply because $\frac{d}{dx}x^p = 0$. As an instance of that, the uniqueness statement derived from (9-7) now holds only up to order q^{p-1}, because of the division by k^m. If one truncates the formula (9-8) modulo q^p, then all terms appearing in it are defined with $\mathbb{F}_p$–coefficients; and it yields the unique solution modulo q^p of (9-4), whose q^0 term equals the cup product with x.

9b Application to quantum Steenrod operations

We adapt our previous definition of quantum Steenrod operations to the weakly monotone context, by adding the variable q. This means that, with (t,θ) as in (1-5) (and omitting the manifold X for the sake of brevity),

$$Q\,\mathrm{St}_p = \sum_A q^{\Omega_X\cdot A}\, Q\,\mathrm{St}_{p,A}\colon H^*(X;\mathbb{F}_p) \to H^*(X;\mathbb{F}_p)[t,\theta][\![q]\!]. \tag{9-12}$$

We find it convenient to introduce a minor generalization, which is a bilinear map on cohomology. More precisely, for each $x \in H^*(X;\mathbb{F}_p)$ one gets an endomorphism of $H^*(X;\mathbb{F}_p)[t,\theta][\![q]\!]$, denoted by

$$Q\Sigma_{p,x} = \sum_A q^{\Omega_X\cdot A} Q\Sigma_{p,x,A}\colon H^*(X;\mathbb{F}_p)[t,\theta][\![q]\!] \to H^*(X;\mathbb{F}_p)[t,\theta][\![q]\!]. \tag{9-13}$$

Geometrically, while quantum Steenrod operations are obtained from holomorphic maps which have (5-19) as a domain, we use the remaining $\mathbb{Z}/p$–fixed point ($z=0$) on that curve as an additional input point to define (9-13). In other words, one can view it as an equivariant version of the "quantum cap product", obtained from the $\mathbb{Z}/p$–equivariant curve in Figure 23. On a technical level, the definition is entirely parallel to that of quantum Steenrod operations, by looking at moduli spaces parametrized by cycles in the classifying space $B\mathbb{Z}/p$ [66]. The q^0 term of (9-13) is the cup product

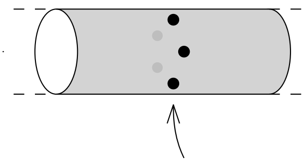

Figure 23: The Riemann surface underlying the definition of $Q\Sigma_{p,x}(\cdot)$; see (9-13).

with the classical Steenrod operation,

$$Q\Sigma_{p,x,0}(y) = \mathrm{St}_p(x)y. \tag{9-14}$$

The relation between (9-12) and (9-13) is that

$$Q\,\mathrm{St}_p(x) = Q\Sigma_{p,x}(1). \tag{9-15}$$

Remark 9.1 It is natural to extend the definition of (9-13) to $x \in H^*(X;\mathbb{F}_p)[\![q]\!]$ in a Frobenius-twisted way, meaning that $\Sigma_{p,qx} = q^p\Sigma_{p,x}$. Then,

$$Q\Sigma_{p,x_1} \circ Q\Sigma_{p,x_2} = (-1)^{p(p-1)/2|x_1||x_2|}Q\Sigma_{p,x_1*x_2}. \tag{9-16}$$

Note that as a consequence of (9-14) and (9-16),

$$\begin{aligned} Q\Sigma_{p,x}(Q\,\mathrm{St}_p(y)) = Q\Sigma_{p,x} \circ Q\Sigma_{p,y}(1) &= (-1)^{p(p-1)/2|x||y|}Q\Sigma_{p,x*y}(1) \\ &= (-1)^{p(p-1)/2|x||y|}Q\,\mathrm{St}_p(x*y). \end{aligned} \tag{9-17}$$

Every class in $H^*(X;\mathbb{F}_p)[t,t^{-1},\theta]$ can be written as $\mathrm{St}_p(y)$ for some y. An analogous statement holds for quantum Steenrod operations, and by combining that with (9-17), one sees that $Q\,\mathrm{St}_p$ actually determines $Q\Sigma_p$.

For our purpose, the key point is the following result:

Theorem 9.2 [66, Theorem 1.4] *For any x, the endomorphism $Q\Sigma_{p,x}$ is covariantly constant for the small quantum connection* (9-3), *meaning that it satisfies* (9-4). (We have tacitly extended the coefficient ring to include θ.)

As a consequence of that and the discussion at the end of Section 9a, $Q\Sigma_{p,x}$ is determined modulo q^p by the classical term (9-15). More explicitly, comparison with (9-8) shows that

$$Q\Sigma_{p,x} = \Phi_{\mathrm{St}_p(x)} \quad \text{modulo } q^p. \tag{9-18}$$

Specializing to (9-14) and using (9-10) leads to:

Corollary 9.3 *The low-degree contributions to the quantum Steenrod operation are*:

$$
\begin{aligned}
(9\text{-}19)\quad & \sum_{\Omega_X\cdot A<p} q^{\Omega_X\cdot A}\int_X y\, Q\operatorname{St}_{p,A}(x) \\
&= \int_X y\operatorname{St}_p(x) - t^{-1}\sum_{0<\Omega_X\cdot A<p} q^{\Omega_X\cdot A}\langle y,(1+t^{-1}\psi)^{-1}\operatorname{St}_p(x)\rangle_{2,A} \\
&\quad + t^{-2}\sum_{0<\Omega_X\cdot A<p} q^{\Omega_X\cdot A}\langle (1-t^{-1}\psi)^{-1} y\operatorname{St}_p(x)\rangle_{1,A} \\
&\quad - t^{-3}\sum_{\substack{\Omega_X\cdot A_0>0\\ \Omega_X\cdot A_1>0\\ \Omega_X\cdot(A_0+A_1)<p}}\sum_k q^{\Omega_X\cdot(A_0+A_1)}\langle y,(1+t^{-1}\psi)^{-1}\operatorname{St}_p(x)\,e_k\rangle_{2,A_0} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \cdot\langle (1-t^{-1}\psi)^{-1}e_k^\vee\rangle_{1,A_1}.
\end{aligned}
$$

Note that even though there are negative powers of t in the formula, we know a priori that none of them can appear in $Q\operatorname{St}_p$, so all terms involving them must cancel.

9c A localization argument

The approach to quantum Steenrod operations via Theorem 9.2 is formally slick, but maybe somewhat indirect; we will therefore suggest a possible alternative. For simplicity, we will work out only the most elementary case. Namely, let's assume that our symplectic manifold X is an algebraic variety, and that we use the given complex structure. We fix some $A\in H_2(X;\mathbb{Z})$ which is holomorphically indecomposable: this means that one can't find nonzero classes $A_1,\dots,A_r$, $r\geq 2$, each of them represented by a holomorphic map $\mathbb{C}P^1\to X$, such that $A_1+\cdots+A_r=A$. This implies that the space of unparametrized rational curves in class A is compact, and contains no multiple covers. We further assume that this space is regular. Consider the standard framework involving stable map spaces,

$$
(9\text{-}20)\qquad
\begin{array}{ccccc}
 & & L_{p+1,0},\dots,L_{p+1,p} & & \\
 & & \downarrow & & \\
\overline{\mathcal{M}}_{p+1} & \longleftarrow & \overline{\mathcal{M}}_{p+1}(X;A) & \xrightarrow{\mathrm{ev}_{p+1}=(\mathrm{ev}_{p+1,0},\dots,\mathrm{ev}_{p+1,p})} & X^{p+1}
\end{array}
$$

Here, $\overline{\mathcal{M}}_{p+1}$ is genus-zero Deligne–Mumford space (we prefer to use the conventional algebrogeometric notation rather than that in the rest of the paper); $\overline{\mathcal{M}}_{p+1}(X;A)$ is the space of stable maps; and the $L_{p+1,k}$ are the tautological line bundles (cotangent bundles of the curve) at the marked points. Our assumption was that $\overline{\mathcal{M}}_0(X;A)$ is regular, hence smooth of complex dimension $n+c_1(A)-3$. As a consequence, all $\overline{\mathcal{M}}_{p+1}(X;A)$ are smooth of dimension $n+c_1(A)+(p-2)$, and actually fiber bundles over $\overline{\mathcal{M}}_0(X;A)$ with fiber $\overline{\mathcal{M}}_{p+1}(\mathbb{C}P^1;1)$. The Sym_p–action on Deligne–Mumford space has a canonical lift to $\overline{\mathcal{M}}_{p+1}(X;A)$.

At this point, we (re)impose the assumption that p is prime. Let $\overline{\mathcal{M}}^{\diamond}_{p+1}(X;A) \subset \overline{\mathcal{M}}_{p+1}(X;A)$ be the subset of stable maps which, under the forgetful map to Deligne–Mumford space, are mapped to the point from (5-19). Clearly, that subset is invariant under $\mathbb{Z}/p \subset \mathrm{Sym}_p$. As before, one can describe its geometry explicitly: $\overline{\mathcal{M}}^{\diamond}_{p+1}(X;A)$ is a fiber bundle over $\overline{\mathcal{M}}_0(X;A)$ with three-dimensional fiber $\overline{\mathcal{M}}^{\diamond}_{p+1}(\mathbb{C}P^1;1)$, and $\mathbb{Z}/p$ acts in a fiber-preserving way.

Lemma 9.4 *The fixed-point set $F \subset \overline{\mathcal{M}}^{\diamond}_{p+1}(X;A)$ of the $\mathbb{Z}/p$–action is the disjoint union of*:

(i) *A copy of $\overline{\mathcal{M}}_2(X;A)$. The restriction of ev_{p+1} to that component can be identified with $(\mathrm{ev}_{2,0}, \mathrm{ev}_{2,1}, \dots, \mathrm{ev}_{2,1})$. Moreover, the normal bundle N of this component is the dual of the tautological bundle $L_{2,1} \to \overline{\mathcal{M}}_2(X;A)$, and the $(\mathbb{Z}/p)$–action on it has weight -1.*

(ii) *A copy of $\overline{\mathcal{M}}_1(X;A)$. The restriction of ev_{p+1} to that component can be identified with $(\mathrm{ev}_{1,0}, \dots, \mathrm{ev}_{1,0})$. Topologically, the normal bundle N of this component is a direct sum of a trivial line bundle and the dual of $L_{1,0}$, with the $(\mathbb{Z}/p)$–action having weight 1 on each component.*

Proof (i) The relevant rational curves have two components $C = C_- \cup_\zeta C_+$; see Figure 24. C_- carries marked points $z_1, \dots, z_p$ and also the node ζ, and can be identified with (5-19), in such a way that ζ corresponds to ∞; the stable map is

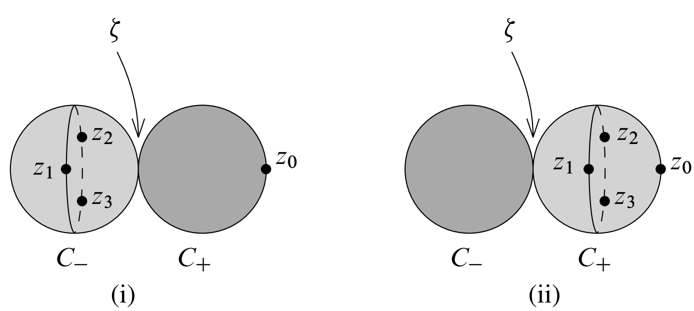

Figure 24: The fixed loci from Lemma 9.4 (with $p = 3$). The lighter shaded components are those where the map is constant.

constant on that component. The other component C_+ carries the node ζ and the marked point z_0; the stable map on that component represents A. The fiber of the normal bundle to the fixed locus at such a point can be canonically identified with $T_\zeta C_- \otimes T_\zeta C_+$; see eg [31, Proposition 3.31]. The identification of C_- with (5-19) mentioned above provides a distinguished isomorphism $T_\zeta C_- \cong \mathbb{C}$ over the entire stratum, and also shows that the action of $\mathbb{Z}/p$ on $T_\zeta C_-$ has weight -1. By definition, $T_\zeta C_+$ is the dual of the cotangent line which is the fiber of $L_{2,1}$, and carries the trivial $\mathbb{Z}/p$–action.

(ii) Here, the curves also have two components $C = C_- \cup_\zeta C_+$, with details as follows; see again Figure 24. There are no marked points on C_-, and $u|C_-$ represents A. The other component is isomorphic to (5-19), compatibly with all marked points and so that the node ζ corresponds to $0 \in \overline{\mathbb{C}}$; and $u|C_+$ is constant. The fiber of the normal bundle to the fixed locus at such a point can be written as an extension

$$0 \to T_\zeta(C_+) \to N \to T_\zeta(C_-) \otimes T_\zeta(C_+) \to 0. \tag{9-21}$$

The tensor product in the right term again expresses gluing together the two components. This time, because ζ is identified with the point 0 in (5-19), the $\mathbb{Z}/p$–action has weight 1 on $T_\zeta(C_+) \cong \mathbb{C}$. The subspace on the left in (9-21) corresponds to staying inside the stratum of nodal curves, but moving the position of the node on C_+. Topologically (9-21) splits, even compatibly with the $(\mathbb{Z}/p)$–action, leading to the desired statement. $\square$

Reformulating the definition of quantum Steenrod operations, one can say that the pairing $(y, x) \mapsto \int_X y Q\, \mathrm{St}_{p,A}(x)$ is obtained as follows:

(9-22)
$$\begin{tikzcd}
H^j(X;\mathbb{F}_p) \otimes H^l(X;\mathbb{F}_p) \arrow[r] \arrow[dd] & H^{pj}_{\mathbb{Z}/p}(X^p;\mathbb{F}_p) \otimes H^l(X;\mathbb{F}_p) \arrow[d, equal] \\
& H^{pj+l}_{\mathbb{Z}/p}(X^{p+1};\mathbb{F}_p) \arrow[d, "\mathrm{ev}^*_{p+1}"'] \\
H^{pj+l-2n-2c_1(A)+6}_{\mathbb{Z}/p}(\mathrm{point};\mathbb{F}_p) & H^{pj+l}_{\mathbb{Z}/p}(\bar{\mathfrak{M}}^\diamond_{p+1}(X;A)) \arrow[l, "\int^{\mathbb{Z}/p}_{\bar{\mathfrak{M}}^\diamond_{p+1}(X;A)}"]
\end{tikzcd}$$

The $\to$ is (the topological version of) the equivariant diagonal map (5-11), and the integration map $\leftarrow$ is the pairing with the equivariant fundamental class of the moduli space. The localization theorem for $\mathbb{Z}/p$–actions [1, Proposition 5.3.18] shows that this integral can be computed in terms of the fixed locus:

$$\int^{\mathbb{Z}/p}_{\bar{\mathcal{M}}^{\diamond}_{p+1}(X;A)} w = \int_F (w|F) e_{\mathbb{Z}/p}(N)^{-1} \quad \text{for } w \in H^*_{\mathbb{Z}/p}(\bar{\mathcal{M}}^{\diamond}_{p+1}(X;A)), \tag{9-23}$$

where $e_{\mathbb{Z}/p}(N)$ is the equivariant Euler class of the normal bundle. Applying this to the description from Lemma 9.4, with w being a class pulled back by evaluation, we get:

Corollary 9.5 *The contribution of a holomorphically indecomposable class A, which has regular moduli spaces, to the quantum Steenrod operation is*

$$\begin{aligned}\int_X y\, Q\,\mathrm{St}_{p,A}(x) \\ = -t^{-1}\langle y, (1+t^{-1}\psi)^{-1}\,\mathrm{St}_p(x)\rangle_{2,A} + t^{-2}\langle (1-t^{-1}\psi)^{-1}\, y\,\mathrm{St}_p(x)\rangle_{1,A}.\end{aligned} \tag{9-24}$$

The use of an integrable complex structure is not really necessary. It was convenient for expository purposes, because it makes the moduli spaces into differentiable manifolds, so that we can talk about the normal bundle to the fixed locus. However, (9-23) also applies to actions on topological manifolds, provided that they are linear in local topological charts near the fixed locus (thereby defining a notion of normal bundle). One can prove that property for regular moduli spaces of pseudoholomorphic curves by standard gluing methods.

Clearly, (9-24) is compatible with the formula (9-19), which we obtained by other means (conversely, one could specialize our earlier argument to only use the class A, and thereby recover (9-24) from it). In principle, this localization method should apply more generally to classes A that are p–indecomposable: by that, we mean that one can't write $A = pA_1 + A_2 + \cdots + A_r$ for $r \geq 1$ and nonzero classes $A_1, \dots, A_r$ which are represented by holomorphic curves (this is always satisfied if $\Omega_X \cdot A < p$). In that situation, there is another component to the fixed locus, which corresponds to the final sum in (9-19). However, note that there are significant technical issues: one can no longer assume that the moduli spaces under consideration are smooth, hence presumably needs to apply a virtual analogue of localization, analogous to [30]. Maybe the most salient argument in favor of the direct approach is that it shows how, when going beyond the situations we have considered so far, the existence of p–fold covered curves complicate the situation: such curves yield yet more components of the fixed locus of the $\mathbb{Z}/p$–action, whose contributions would need to be studied separately.

9d Basic examples

From this point onwards, we return to the monotone context from the main part of the paper — actually, our first two examples are only spherically monotone, meaning that the symplectic class and first Chern class are positive proportional on $\pi_2(X)$, but that's just as good for our purpose. In terms of formulae such as (9-19), this simply means that all expressions are polynomials in q for degree reasons, and hence it is permitted to remove the formal variable by setting $q = 1$.

Example 9.6 Let $X = T^2 \times \mathbb{C}P^2$. We consider $p = 2$, take any $x \in H^1(T^2; \mathbb{F}_2)$, and let $l \in H^2(\mathbb{C}P^2; \mathbb{F}_2)$ be the generator. The classical contribution is

$$\begin{cases} \Xi_2(x \otimes 1) = x \otimes 1, \\ \Xi_2(x \otimes l) = x \otimes l^2, \\ \Xi_2(x \otimes l^2) = 0. \end{cases} \tag{9-25}$$

For degree reasons, the only other contribution comes from the class A of a line in $\mathbb{C}P^2$, and in view of of invariance under symplectic automorphisms, that contribution must be that $Q\Xi_{2,A}(x \otimes l^2)$ is some multiple of $x \otimes l$. Using the notation $\theta = t^{1/2}$, we have $Q\,\mathrm{St}(x \otimes l^2) = t^{5/2}(x \otimes l^2)$. Take some $y \in H^1(T^2; \mathbb{F}_2)$. Then (9-24) says that

$$\begin{aligned} \int_X (y \otimes l)\, Q\,\mathrm{St}_{2,A}(t^{5/2}(x \otimes l^2)) &= t^{1/2} \langle y \otimes l, \psi\,(x \otimes l^2)\rangle_{2,A} \\ &= t^{1/2}\Bigl(\int_{T^2} yx\Bigr)\langle l, \psi l^2\rangle_{\mathbb{C}P^2,2,A}. \end{aligned} \tag{9-26}$$

In the rightmost term, the Gromov–Witten invariant is taken in $\mathbb{C}P^2$, and we have marked that notationally (to get to that expression, we have used the fact that all our curves are constant in T^2–direction). Geometrically, what we are considering in that term is the space of all lines in $\mathbb{C}P^2$ going through a specific line Z (representing l) and a point $q \notin Z$ (representing l^2). That moduli space can be identified with Z, and the normal bundle to Z is the dual to the line bundle which gives rise to the gravitational descendant in the formula. Hence, $\langle l, \psi l^2\rangle_{\mathbb{C}P^2,2,A} = 1$.

Let's pass to the algebraic closure $\overline{\mathbb{F}}_2$, with a nontrivial third root of unity $\zeta \in \overline{\mathbb{F}}_2$. This yields a splitting of $1 \in QH^*(\mathbb{C}P^2; \overline{\mathbb{F}}_2)$ into idempotents $u_j = 1 + \zeta^j l + \zeta^{2j} l^2$. The natural extension of $Q\Xi_2$ to $H^{\mathrm{odd}}(X; \overline{\mathbb{F}}_2)$ is linear in a Frobenius-twisted sense, meaning that $Q\Xi_2(\zeta\,\cdot) = \zeta^2 Q\Xi_2(\cdot)$. From our previous computations, it follows that

$$Q\Xi_2(x \otimes u_j) = x \otimes u_j. \tag{9-27}$$

Let's see how this might look from a categorical viewpoint, along the lines of Remark 1.10. The Fukaya category of $\mathbb{C}P^2$ over $\overline{\mathbb{F}}_2$ is more properly described as a collection of three categories, each of which is semisimple, and which correspond to the idempotent summands $\overline{\mathbb{F}}_2 u_j$ in quantum cohomology. Unfortunately, we cannot introduce a meaningful version of the Fukaya category of T^2 without Novikov parameters, because there are no nontrivial monotone Lagrangian submanifolds. However, the diagonal in the Fukaya category of $X \times X$ makes sense in a monotone context, so we can frame the discussion in that language. Algebraically, the diagonal splits as the direct sum of three objects, again corresponding to the u_j. We can take one of those summands and equip it with a flat line bundle, or formal family of flat line bundles, in the T^2–direction. The convolution of such Lagrangian correspondences gives us $H^1(T^2; \mathbb{G}_m) = \mathbb{G}_m^2$, respectively its formal completion. Since we can do that independently for all three summands, we see the formal completion of $\mathbb{G}_m^6$ appearing, which is consistent with (9-27). If one wanted to work over $\mathbb{F}_2$ itself, only the idempotents u_0 and $u_1 + u_2$ would be defined, giving rise to a more complicated picture of the Fukaya category.

Example 9.7 Along the same lines, take $X = T^2 \times \mathbb{C}P^1 \times \mathbb{C}P^1$, but now with $p = 3$. Take $x, y \in H^1(T^2; \mathbb{F}_3)$ and $k, l \in H^2(\mathbb{C}P^1 \times \mathbb{C}P^1; \mathbb{F}_3)$. The only classes A which contribute to $Q\Xi_3(x \otimes k)$ are those which yield the two rulings of X, and hence satisfy $\langle[\text{point}]\rangle_{1,A} = 1$. For each such class, (9-24) yields

$$\int_X (y \otimes l) Q\Xi_{3,A}(x \otimes k) = -\langle (y \otimes l)(x \otimes k)\rangle_{1,A} = -\int_X (y \otimes l)(x \otimes k). \tag{9-28}$$

Adding up their contributions yields

$$Q\Xi_3 = -2\,\mathrm{id} = \mathrm{id} \quad \text{on } H^3(X; \mathbb{F}_3). \tag{9-29}$$

The same holds on $H^1(X; \mathbb{F}_3)$; see (1-14). The corresponding question for $H^5(X; \mathbb{F}_3)$ is just outside the reach of our methods, because there is a potential contribution from classes that are 3 times that of a ruling.

This time, the Fukaya category of $\mathbb{C}P^1 \times \mathbb{C}P^1$ splits into four semisimple pieces over $\mathbb{F}_3$, so one expects to see the product of four copies of the formal group associated to T^2, meaning a total of $\widehat{\mathbb{G}}_m^8$, which is compatible with our (partial) computation.

9e Fano threefolds

The remaining examples will be monotone symplectic six-manifolds, which have $H_1(X; \mathbb{Z}) = 0$ and $H_*(X; \mathbb{Z})$ torsion-free (in fact, they will be algebraic, meaning

Fano threefolds). We will assume that there is some $\lambda \in \mathbb{Z}$ such that

$$c_1(X) * x = \lambda x \quad \text{for } x \in H^3(X), \tag{9-30}$$

or, equivalently,

$$\langle y, x\rangle_2 = \lambda \int_X yx \quad \text{for } x, y \in H^3(X).$$

From now on, we fix a prime p, and our notation will be that $x, y \in H^3(X;\mathbb{F}_p)$. The classical Steenrod operations applied to x have potentially nontrivial components in degrees 3, 4 and 6. The degree 4 component is the Bockstein β, which is zero because all our classes come from $H^3(X;\mathbb{Z})$. The degree 6 component is the $t^{(3p-7)/2}\theta$ part of $\mathrm{Sq}(x)$. For $p = 2$, this is just the cup square, which again is zero by lifting to $H^3(X;\mathbb{Z})$; and for $p > 2$, it involves the Bockstein, see (1-6), hence is again zero. The outcome is that only the degree 3 component survives. Taking the constants in (1-6) into account (and omitting X from the notation), this says that

$$\mathrm{St}_p(x) = \begin{cases} xt^{3/2} & \text{if } p = 2, \\ -\left(\frac{p-1}{2}!\right)xt^{(3p-3)/2} & \text{if } p > 2. \end{cases} \tag{9-31}$$

As a consequence of (9-30) (and the divisor and TRR relations in Gromov–Witten theory), we have for $0 < d \le p-2$,

$$\langle y, \psi^d x\rangle_2 = \frac{\lambda}{d+1}\langle y, \psi^{d-1}x\rangle_2 = \cdots = \frac{\lambda^{d+1}}{(d+1)!}\int_X yx. \tag{9-32}$$

Let's look at $\int_X y\, Q\Xi_p(x)$. For degree reasons, the curves that contribute to this lie in classes A with $c_1(A) = p-1$, hence (setting $\Omega_X = c_1(X)$ in view of monotonicity) Corollary 9.3 applies. At first sight, the outcome reads as follows:

$$\begin{aligned} &\int_X yQ\Xi_p(x) \\ &\quad = -\sum_{\Omega\cdot A = p-1} \langle y, \psi^{p-2}x\rangle_{2,A} - \int_X yx \sum_{\Omega\cdot A=p-1} \langle \psi^{p-3}[\text{point}]\rangle_{1,A} \\ &\qquad + \sum_{\substack{d_0=\Omega\cdot A_0>0\\ d_1=\Omega\cdot A_1>0\\ d_0+d_1=p-1}} \langle y, (-1)^{d_0-1}\psi^{d_0-1}x\rangle_{2,A_0}\langle\psi^{d_1-2}[\text{point}]\rangle_{1,A_1}. \end{aligned} \tag{9-33}$$

Using (9-32), one simplifies this to

$$\begin{aligned} &\int_X y\, Q\Xi_p(x) \\ &\quad = \left(\int_X xy\right)\left(-\frac{\lambda^{p-1}}{(p-1)!} + \sum_{2\le d\le p-1} (-1)^{d-1}\frac{\lambda^{p-1-d}}{(p-1-d)!}\langle\psi^{d-2}[\text{point}]\rangle_1\right). \end{aligned} \tag{9-34}$$

Besides λ, the enumerative ingredient that enters is the quantum period (see [12], where the notation is G_X)

$$\Pi = 1 + \sum_{d \geq 2} q^d \langle \psi^{d-2}[\mathrm{point}] \rangle_{1,A}, \tag{9-35}$$

or more precisely, what's obtained from it by truncating mod q^p and then considering the coefficients as lying in $\mathbb{F}_p$. In that notation, one can also write (9-34) as

$$Q\Xi_p = -(q^{p-1}\text{–coefficient of } e^{-\lambda q}\,\Pi)\,\mathrm{id}. \tag{9-36}$$

All the examples that we will consider are instances of [12, Theorem 4.7], itself based on Givental's work.

Example 9.8 Let X be the intersection of two quadrics in $\mathbb{C}P^5$, which is also a moduli space of stable bundles (with rank two and fixed odd-degree determinant) on a genus two curve. This has

$$H_l(X;\mathbb{Z}) = \begin{cases} \mathbb{Z}^4 & \text{if } l = 3, \\ \mathbb{Z} & \text{if } l = 0, 2, 4, 6, \\ 0 & \text{otherwise.} \end{cases} \tag{9-37}$$

The first Chern class is twice a generator of $H^2(X;\mathbb{Z})$. For degree reasons, this implies that $Q\Xi_2 = 0$. This is not necessarily indicative of the general picture, since we already know that the prime $p = 2$ is exceptional [15, page 137]: the small quantum cohomology ring has $QH^{\mathrm{even}}(X) \cong \mathbb{Z}[h]/h^2(h^2 - 16)$, hence does not split into summands if one reduces coefficients to $\mathbb{F}_2$.

Let's look at odd primes. We have $\lambda = 0$ since there are no classes A with $\Omega_X \cdot A = 1$. The quantum period is [13, page 135]

$$\Pi = \sum_{d \geq 0} \frac{(2d)!^2}{(d!)^6} q^{2d} = 1 + 2^2 q^2 + 3^2 q^4 + \left(\tfrac{10}{3}\right)^2 q^6 + \left(\tfrac{35}{12}\right)^2 q^8 + \cdots. \tag{9-38}$$

Applying (9-34) yields

$$Q\Xi_p = -\frac{(p-1)!^2}{\left(\frac{1}{2}(p-1)\right)!^6}\,\mathrm{id} = (-1)^{(p-1)/2}\,\mathrm{id} \quad \text{for odd } p. \tag{9-39}$$

We should point out that the first nontrivial case $p = 3$, where the enumerative geometry is that of lines on X, is amenable to the more direct method of Section 9c. The space of lines is regular [59, Theorem 2.6] (it is isomorphic to the Jacobian of the genus two curve associated to X [59, Theorem 4.7]), and there are 4 lines passing through a generic point [16, page 135]. That information enables one to apply (9-24) and obtain $Q\Xi_3 = -4\,\mathrm{id} = -\mathrm{id}$.

When thinking about the outcome of this computation, it may be useful to know that there is an algebraic group whose formal completion shows the same behavior, namely

$$G = \{x + iy : x^2 + y^2 = 1\}, \tag{9-40}$$

where i is an abstract symbol such that $i^2 = -1$ (some readers may feel more comfortable writing this as a group of 2×2 rotation matrices). It is a nonsplit torus, which becomes isomorphic to $\mathbb{G}_m$ over any coefficient ring that contains an actual root of -1. To write down the group law for the completion $\widehat{G}$, one can use the rational parametrization $z = y/(1+x)$, in which it is given by

$$z_1 \bullet z_2 = \frac{z_1 + z_2}{1 - z_1 z_2}. \tag{9-41}$$

The p^{th} power map, for primes $p > 2$, is $(x+iy)^p \equiv x^p + (-1)^{(p-1)/2} i y^p \bmod p$, or for (9-41),

$$\overbrace{z \bullet \cdots \bullet z}^{p} \equiv \frac{(-1)^{(p-1)/2} y^p}{1 + x^p} = (-1)^{(p-1)/2} z^p = (-1)^{(p-1)/2} z \quad \text{for } z \in \mathbb{F}_p. \tag{9-42}$$

A natural conjecture would be that the formal group associated to X is $\widehat{G}^4$. Note that in [69], a direct summand of the Fukaya category of X was shown to be equivalent to the Fukaya category of the genus two curve. This seems to suggest a role for $\widehat{\mathbb{G}}_m^4$ rather than $\widehat{G}^4$ (compare Remark 1.10). However, [69] works with complex number coefficients. To the author's best knowledge, we do not have a version of that argument that would work over $\mathbb{Z}$ or $\mathbb{F}_p$, and hence, it remains an open question to interpret the computation above in terms of mirror symmetry.

Example 9.9 Let $X \subset \mathbb{C}P^4$ be a cubic threefold. This situation is parallel to the previous example, except that $H^3(X)$ is ten-dimensional. The quantum period is [13, page 134]

$$\Pi = \sum_d \frac{(3d)!}{(d!)^5} q^{2d} = 1 + 6q^2 + \tfrac{45}{2} q^4 + \tfrac{140}{3} q^6 + \tfrac{1925}{32} q^8 + \cdots. \tag{9-43}$$

One again has $Q\Xi_2 = 0$ and $\lambda = 0$ for degree reasons, but this time

$$Q\Xi_p = -\frac{\left(\frac{3p-3}{2}!\right)}{\left(\frac{p-1}{2}!\right)^5} \,\mathrm{id} = 0 \quad \text{for } p > 2. \tag{9-44}$$

Example 9.10 The quartic threefold has [13, page 136]

$$\Pi = e^{-24q} \sum_d \frac{(4d)!}{(d!)^5} q^d. \tag{9-45}$$

We have $\lambda = -24$, the "big eigenvalue" in the terminology of [67, Corollary 1.14]. The e^{-24q} cancels out the corresponding term in (9-36), and we again have $Q\Xi_p = 0$.

In both of the previous examples, homological mirror symmetry is known to hold [67], but again, arithmetic aspects are not addressed in that paper. Even assuming that the answer given there is true arithmetically, it takes the form of an orbifold LG (Landau–Ginzburg) model with a highly degenerate singular point. For our purpose, one would need to understand the formal completion near the identity of the derived automorphism group of such an LG model, which is a purely algebrogeometric problem, but one whose answer is not known to this author. This means that we do not have a mirror symmetry interpretation for the vanishing of $Q\Xi_p$.

Example 9.11 (previously mentioned in Example 1.11) Let X be a hypersurface of bidegree $(1,2)$ in $\mathbb{C}P^1 \times \mathbb{C}P^3$; equivalently, this is obtained by blowing up the intersection of two quadrics (which is an elliptic curve) in $\mathbb{C}P^3$. It is a Fano threefold satisfying

$$H_l(X;\mathbb{Z}) = \begin{cases} \mathbb{Z}^2 & \text{if } l = 2,3,4, \\ \mathbb{Z} & \text{if } l = 0,6, \\ 0 & \text{otherwise.} \end{cases} \tag{9-46}$$

More precisely, we have $H_2(X;\mathbb{Z}) \cong H_2(\mathbb{C}P^1 \times \mathbb{C}P^3;\mathbb{Z})$ by inclusion, and for the exceptional divisor $T^2 \times \mathbb{C}P^1 \subset X$, we similarly have $H_3(T^2 \times \mathbb{C}P^1) \cong H_3(X;\mathbb{Z})$. The classes potentially represented by holomorphic curves are

$$A = (d_1, d_2), \quad \text{with } d_1, d_2 \geq 0 \text{ and } \Omega_X \cdot A = d_1 + 2d_2. \tag{9-47}$$

Curves in the unique class $A = (1,0)$ with $\Omega_X \cdot A = (1,0)$ form the ruling of the exceptional divisor. From that, one easily sees that $\lambda = -1$. We have [13, page 183]

$$e^{-\lambda q}\Pi = e^q \Pi = \sum_{d_1,d_2} \frac{(d_1 + 2d_2)!}{(d_1!)^2 (d_2!)^4} q^{d_1+2d_2}, \tag{9-48}$$

and hence

$$Q\Xi_p = \mathrm{id} \sum_{d_1+2d_2=p-1} \frac{1}{(d_1!)^2(d_2!)^4}. \tag{9-49}$$

See (1-16) for the first few terms. As before, the lowest-degree case $p = 2$ is amenable to more direct methods, and was in fact determined in [74].

One can write

$$\text{(9-50)}\quad \sum_{d_1+2d_2=m} \frac{m!^2}{(d_1!)^2(d_2!)^4} = \sum_{d_1+2d_2=m} \binom{m}{d_1 d_2 d_2}^2 = \text{constant coefficient of } \widetilde{W}^m,$$

where

$$\widetilde{W}(x_0,x_1,x_2,x_3) = \frac{(x_0^2+x_1^2+x_2x_3)(x_0x_1+x_2^2+x_3^2)}{x_0x_1x_2x_3}.$$

This is an elementary combinatorial argument: when we expand

$$(x_0^2+x_1^2+x_2x_3)^m(x_0x_1+x_2^2+x_3^2)^m,$$

the monomial $(x_0x_1x_2x_3)^m$ arises by picking each x_k^2 term an equal number (d_2 in our formula) of times, which leads to the multinomial coefficients. Setting $m = p-1$ allows us to apply that to (9-49). Next, consider the intersection of the two quadrics that appear in (9-50),

$$\text{(9-51)}\qquad C = \{x_0^2+x_1^2+x_2x_3 = 0,\ x_0x_1+x_2^2+x_3^2 = 0\} \subset \mathbb{P}^3.$$

There is an elementary number theory argument which allows one to count the number of points of $C(\mathbb{F}_p)$ modulo p; it goes as follows. By little Fermat,

$$\text{(9-52)}\qquad \sum_{x_0,\dots,x_3} (1-x_0^2-x_1^2-x_2x_3)^{p-1}(1-x_0x_1-x_2^2-x_3^2)^{p-1} \in \mathbb{F}_p$$

counts the number of points in $\mathbb{F}_p^4$ lying on the intersection of our quadrics. On the other hand,

$$\text{(9-53)}\qquad \begin{aligned}\sum_{x_0,\dots,x_3} x_0^{i_0}\cdots x_3^{i_3} &= \Bigl(\sum_{x_0} x_0^{i_0}\Bigr)\Bigl(\sum_{x_1} x_1^{i_1}\Bigr)\Bigl(\sum_{x_2} x_2^{i_2}\Bigr)\Bigl(\sum_{x_3} x_3^{i_3}\Bigr)\\ &= \begin{cases} 1 & \text{for } (i_0,\dots,i_3) = (p-1,\dots,p-1),\\ 0 & \text{for all other } 0 \le i_0,\dots,i_3 \le p-1.\end{cases}\end{aligned}$$

If we expand (9-52) and apply (9-53) to the resulting terms, the outcome is that (9-52) is the $x_0^{p-1}x_1^{p-1}x_2^{p-1}x_3^{p-1}$–coefficient of $(x_0^2-x_1^2-x_2x_3)^{p-1}(x_0x_1-x_2^2-x_3^2)^{p-1}$, which is what appears in (9-50). Adjusting that to the point-count in projective space, we get

$$\text{(9-54)}\qquad 1-\#C(\mathbb{F}_p) \equiv \text{constant coefficient of } \widetilde{W}^m \mod p.$$

The isogeny class of the elliptic curve $\mathrm{Jac}(C)$ is listed as [43, Isogeny class 15.a], and its associated modular form is (1-17). Point-counting becomes relevant for us through a theorem of Honda [34; 56; 32], which says that $1-\#C(\mathbb{F}_p) \bmod p$ can be identified with the p^{th} coefficient of the p^{th} power map for the formal group which is the completion of $\mathrm{Jac}(C)$ (this coefficient is sometimes called the Hasse invariant of the mod p reduction of C; maybe more precisely, it is a special case of the Hasse–Witt matrix of an algebraic curve). The natural interpretation of this in terms of mirror symmetry is the following:

Conjecture 9.12 Take X as in Examples 1.11 and 9.11. The Fukaya category of X contains a direct summand equivalent to the derived category of sheaves on a genus one curve, whose Jacobian is isogeneous to that of C. (The remaining summands are expected to be semisimple: they do not contribute to $H^{\mathrm{odd}}(X)$ or to our formal group.)

It turns out that this is compatible with predictions coming from "classical" enumerative mirror symmetry. There (see eg [12, Definition 4.9]), a mirror superpotential $W \in \mathbb{Z}[y_1^{\pm 1}, y_2^{\pm 1}, y_3^{\pm 1}]$ for X needs to have the property that

$$\begin{aligned}\Pi &= \int_{|y_1|=|y_2|=|y_3|=1} e^{qW}\frac{dy_1\wedge dy_2\wedge dy_3}{y_1y_2y_3} \\ &= \sum_{d=0}^{\infty}\frac{\text{constant coefficient of } W^d}{d!}q^d.\end{aligned} \tag{9-55}$$

There can be infinitely many different superpotentials for the same X, related by certain birational changes of variables. Assuming that the anticanonical linear system for X contains a smooth divisor, then the actual mirror of X, formed relative to that divisor, should come with a proper (fibers are compact) function that specializes to those superpotentials in different Zariski charts.

Getting back to our example: the function $\widetilde{W}(1,x_1,x_2,x_3)-1$ satisfies the property (9-55), as a consequence of (9-50), but fails another requirement for mirror superpotentials, that of having a reflexive Newton polyhedron. Instead, the precise relation is as follows. One of the superpotentials for our specific X, given in [11, Polytope 198], is

$$W = y_1+y_2+y_3+y_1^{-1}y_2y_3+y_1y_3^{-1}+y_2^{-1}+y_1^{-1}+y_2^{-1}y_3^{-1}. \tag{9-56}$$

One can then write

$$\widetilde{W}(1,x_1,x_2,x_3)-1 = W(x_1^{-1}x_2x_3^{-1}, x_1^{-1}x_2^{-1}x_3, x_1^{-1}x_2^2). \tag{9-57}$$

The monomial coordinate change in (9-57) is a $\mathbb{Z}/4$–cover of the (y_1,y_2,y_3)–torus by the (x_1,x_2,x_3)–torus; such coordinate changes do not affect oscillating integrals as in (9-55). It is clear from the definition (9-50) that the critical locus of $\widetilde{W}(1,x_1,x_2,x_3)$

contains an affine part of C, lying in the fiber $\widetilde{W}^{-1}(0)$. Hence, the critical locus of W contains an affine part of a $\mathbb{Z}/4$–quotient of C, lying in the fiber $W^{-1}(-1)$. Moreover, the Hessian in transverse direction to those critical loci is nondegenerate over $\mathbb{Q}$ (or over $\mathbb{F}_p$, provided that p is large). In view of the expected correspondence between the Fukaya category of X and the category D^b_{sing} associated to a compactification of W, this provides strong support for Conjecture 9.12, and also gives a specific candidate genus-one curve (within the given isogeny class).

10 Sign conventions

Signs are important for some of our example computations. This section clarifies the conventions used for $\mathbb{Z}/p$–equivariant (and therefore Sym_p–equivariant) cohomology, and for the Steenrod operations.

10a Equivariant cohomology

Take the standard classifying space $BS^1 = S^\infty/S^1 = \mathbb{C}P^\infty$. Let $t \in H^2_{S^1}(\text{point}) = H^2(\mathbb{C}P^\infty)$ be the Chern class of $\mathcal{O}(-1)$. Given a representation V of S^1, form the associated vector bundle

$$(V \times S^\infty)/S^1 \to \mathbb{C}P^\infty, \quad \text{where } g\cdot(v,z) = (gv, g^{-1}z). \tag{10-1}$$

In this way, the representation V_k of weight k corresponds to the line bundle $\mathcal{O}(-k)$. We use the same convention as in (10-1) when forming the Borel construction (the equivariant cohomology of a space), and similarly for equivariant Euler classes of vector bundles.

Given a representation $V_{k_1} \oplus \cdots \oplus V_{k_d}$, its equivariant Euler class, defined as the Euler class of the associated bundle (10-1), is therefore

$$e_{S^1}(V) = k_1 \cdots k_d\, t^d. \tag{10-2}$$

We embed $\mathbb{Z}/p \subset S^1$ in the obvious way, and take the mod p reduction of t to be the generator of $H^2_{\mathbb{Z}/p}(\text{point}; \mathbb{F}_p)$, leading to a corresponding version of (10-2). We take $\theta \in H^1_{\mathbb{Z}/p}(\text{point}; \mathbb{F}_p)$ to be the tautological generator, meaning the one associated to the identity map $\mathbb{Z}/p = \pi_1(B(\mathbb{Z}/p)) \to \mathbb{F}_p$. Then, the Bockstein satisfies

$$\beta(\theta) = -t. \tag{10-3}$$

Example 10.1 Fix some odd p. Take the fundamental representation of $\mathbb{Z}/p$ on $\mathbb{R}^p$, by cyclic permutations, and let V be its quotient by the trivial subspace $\mathbb{R}(1,\dots,1)$. Our orientation convention is that taking first $(1,\dots,1)$, and then after that lifts of an oriented basis of V, yields an oriented basis of $\mathbb{R}^p$. If we temporarily ignore orientations, then clearly

$$V \cong V_1 \oplus V_2 \oplus \cdots V_{(p-1)/2}. \tag{10-4}$$

This decomposition can be made explicit in terms of a discrete Fourier basis. A computation of the determinant of that basis (compare eg [52]) shows that (10-4) is in fact orientation-preserving. As a consequence,

$$e_{\mathbb{Z}/p}(V) = \left(\frac{p-1}{2}\right)!\, t^{(p-1)/2}. \tag{10-5}$$

Example 10.2 To check the sign in (10-3), let's replace the infinite-dimensional space $K(\mathbb{Z}/p,1)$ by the lens space $L(p,1) = S^3/(\mathbb{Z}/p)$, with $\mathbb{Z}/p$ acting diagonally on $S^3 = \{|z_1|^2+|z_2|^2 = 1\} \subset \mathbb{C}^2$. The relation between the homological Bockstein b and its cohomological counterpart β is that

$$\langle \beta(y), x\rangle + (-1)^{|y|}\langle y, b(x)\rangle = 0. \tag{10-6}$$

Consider $\{|z_1| \leq 1,\ z_2 = \sqrt{1-|z_1|^2}\} \subset L(p,1)$, with the complex orientation from z_1. This is a $\mathbb{Z}/p$–cycle, whose homology class we write as x. Applying the homological Bockstein yields a $1/p$ fraction of the boundary, which is exactly the circle $\{z_2 = 0\} \subset L(p,1)$, with its orientation given by going around z_1 anticlockwise from 1 to $e^{2\pi i/p}$. This means that by definition of θ,

$$\langle \theta, b(x)\rangle = 1. \tag{10-7}$$

The class $-t$ is Poincaré dual to the zero-locus of a section of the pullback of $\mathcal{O}(1)$, hence represented by the cycle $\{z_1 = 0\}$, with the usual orientation of the z_2 circle. The intersection number of that and the mod p cycle defined above is

$$\langle -t, x\rangle = 1. \tag{10-8}$$

From (10-6), for $y = \theta$, and (10-7) we get $\langle \beta(\theta), x\rangle = 1$, which together with (10-8) yields the desired (10-3).

10b Steenrod operations

The appearance of combinatorial constants similar to those in (1-6) goes back to the classical literature; see eg [70, pages 107 and 112]. The point of introducing those is to make sure the operations satisfy the Steenrod axioms. Since a comparison between different definitions is made more complicated by sign conventions for equivariant cohomology, we want to explain one way of checking the choices made here.

Fix an odd prime p. Consider the Steenrod axiom which says that $P^0(x) = x$ for $x \in H^*(X; \mathbb{F}_p)$. With our convention (1-6), this is equivalent to

$$\mathrm{St}_p(x) = (-1)^* \left(\frac{p-1}{2}!\right)^{|x|} t^{(p-1)/2|x|} x + \text{terms of higher degree in } H^*(X; \mathbb{F}_p), \tag{10-9}$$

with $*$ as in (1-7). Suppose that X is an oriented closed manifold, and that we apply this to $x = [\text{point}] \in H^{\dim(X)}(X; \mathbb{F}_p)$. By definition, $\mathrm{St}_p(x)$ is obtained from

$$H^{\dim(X)}(X; \mathbb{F}_p) \xrightarrow{p^{\text{th}} \text{ power}} H^{p\dim(X)}_{\mathbb{Z}/p}(X^p; \mathbb{F}_p) \xrightarrow[\text{}]{\substack{\text{restriction} \\ \text{to diagonal}}} H^{p\dim(X)}_{\mathbb{Z}/p}(X; \mathbb{F}_p). \tag{10-10}$$

Hence, it maps x to itself times the equivariant Euler class of the normal bundle to the diagonal $X \subset X^p$, restricted to a point. If X is one-dimensional, that normal bundle is given by the representation V from Example 10.1. In general, it can be identified with $\dim(X)$ copies of V, up to a Koszul reordering sign $(-1)^\dagger$,

$$\dagger = \frac{|x|(|x|-1)}{2}\frac{p(p-1)}{2} \equiv \frac{|x|(|x|-1)}{2}\frac{p-1}{2} \mod 2. \tag{10-11}$$

Combining that with the $|x|^{\text{th}}$ power of (10-5) precisely yields the constant factor in (10-9).

References

[1] **C Allday**, **V Puppe**, *Cohomological methods in transformation groups*, Cambridge Studies in Advanced Mathematics 32, Cambridge Univ. Press (1993) MR Zbl

[2] **G Alston**, **L Amorim**, *Floer cohomology of torus fibers and real Lagrangians in Fano toric manifolds*, Int. Math. Res. Not. 2012 (2012) 2751–2793 MR Zbl

[3] **S Axelrod**, **I M Singer**, *Chern–Simons perturbation theory, II*, J. Differential Geom. 39 (1994) 173–213 MR Zbl

[4] **Y Bespalov**, **V Lyubashenko**, **O Manzyuk**, *Pretriangulated A_∞–categories*, Pr. Inst. Mat. Nats. Akad. Nauk Ukr. Mat. Zastos. 76, NAS Ukraine, Institute of Math., Kiev (2008) MR Zbl

[5] **N Bottman**, *2–associahedra*, Algebr. Geom. Topol. 19 (2019) 743–806 MR Zbl

[6] **N Bottman**, *Moduli spaces of witch curves topologically realize the 2–associahedra*, J. Symplectic Geom. 17 (2019) 1649–1682 MR Zbl

[7] **N Bottman**, **S Carmeli**, *$(A_\infty, 2)$–categories and relative 2–operads*, High. Struct. 5 (2021) 401–421 MR Zbl

[8] **N Bottman**, **A Oblomkov**, *A compactification of the moduli space of marked vertical lines in $\mathbb{C}^2$*, preprint (2019) arXiv 1910.02037

[9] **N Bottman**, **K Wehrheim**, *Gromov compactness for squiggly strip shrinking in pseudoholomorphic quilts*, Selecta Math. 24 (2018) 3381–3443 MR Zbl

[10] **K S Brown**, *Cohomology of groups*, Graduate Texts in Math. 87, Springer (1982) MR Zbl

[11] **T Coates**, **A Corti**, **S Galkin**, **V Golyshev**, **A Kasprzyk**, *Fano varieties and extremal Laurent polynomials*, electronic reference (2012) Available at `www.fanosearch.net`

[12] **T Coates**, **A Corti**, **S Galkin**, **V Golyshev**, **A Kasprzyk**, *Mirror symmetry and Fano manifolds*, from "European Congress of Mathematics" (R Latała, A Ruciński, P Strzelecki, J Świątkowski, D Wrzosek, P Zakrzewski, editors), Eur. Math. Soc., Zürich (2013) 285–300 MR Zbl

[13] **T Coates**, **A Corti**, **S Galkin**, **A Kasprzyk**, *Quantum periods for 3–dimensional Fano manifolds*, Geom. Topol. 20 (2016) 103–256 MR Zbl

[14] **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

[15] **S K Donaldson**, *Gluing techniques in the cohomology of moduli spaces*, from "Topological methods in modern mathematics" (L R Goldberg, A V Phillips, editors), Publish or Perish, Houston, TX (1993) 137–170 MR Zbl

[16] **S K Donaldson**, *Floer homology and algebraic geometry*, from "Vector bundles in algebraic geometry" (N J Hitchin, P E Newstead, W M Oxbury, editors), London Math. Soc. Lecture Note Ser. 208, Cambridge Univ. Press (1995) 119–138 MR Zbl

[17] **G C Drummond-Cole**, *Homotopically trivializing the circle in the framed little disks*, J. Topol. 7 (2014) 641–676 MR Zbl

[18] **J D Evans**, **Y Lekili**, *Floer cohomology of the Chiang Lagrangian*, Selecta Math. 21 (2015) 1361–1404 MR Zbl

[19] **J Fine**, **D Panov**, *Hyperbolic geometry and non-Kähler manifolds with trivial canonical bundle*, Geom. Topol. 14 (2010) 1723–1763 MR Zbl

[20] **A Floer**, **H Hofer**, **D Salamon**, *Transversality in elliptic Morse theory for the symplectic action*, Duke Math. J. 80 (1995) 251–292 MR Zbl

[21] **K Fukaya**, *Morse homotopy and its quantization*, from "Geometric topology" (W H Kazez, editor), AMS/IP Stud. Adv. Math. 2, Amer. Math. Soc., Providence, RI (1997) 409–440 MR Zbl

[22] **K Fukaya**, *Unobstructed immersed Lagrangian correspondence and filtered A_∞ functor*, preprint (2017) arXiv 1706.02131

[23] **K Fukaya**, **Y-G Oh**, **H Ohta**, **K Ono**, *Lagrangian intersection Floer theory: anomaly and obstruction, I*, AMS/IP Stud. Adv. Math. 46.1, Amer. Math. Soc., Providence, RI (2009) MR Zbl

[24] **K Fukaya**, **Y-G Oh**, **H Ohta**, **K Ono**, *Lagrangian intersection Floer theory: anomaly and obstruction, II*, AMS/IP Stud. Adv Math. 46.2, Amer. Math. Soc., Providence, RI (2009) MR Zbl

[25] **W Fulton**, **R MacPherson**, *A compactification of configuration spaces*, Ann. of Math. 139 (1994) 183–225 MR Zbl

[26] **S Ganatra**, *Symplectic cohomology and duality for the wrapped Fukaya category*, preprint (2013) arXiv 1304.7312

[27] **M Gerstenhaber**, **A A Voronov**, *Homotopy G–algebras and moduli space operad*, Int. Math. Res. Notices (1995) 141–153 MR Zbl

[28] **E Getzler**, *Batalin–Vilkovisky algebras and two-dimensional topological field theories*, Comm. Math. Phys. 159 (1994) 265–285 MR Zbl

[29] **E Getzler**, **J D S Jones**, *Operads, homotopy algebra and iterated integrals for double loop spaces*, preprint (1994) arXiv hep-th/9403055

[30] **T Graber**, **R Pandharipande**, *Localization of virtual classes*, Invent. Math. 135 (1999) 487–518 MR Zbl

[31] **J Harris**, **I Morrison**, *Moduli of curves*, Graduate Texts in Math. 187, Springer (1998) MR Zbl

[32] **M Hazewinkel**, *Three lectures on formal groups*, from "Lie algebras and related topics" (D J Britten, F W Lemire, R V Moody, editors), CMS Conf. Proc. 5, Amer. Math. Soc., Providence, RI (1986) 51–67 MR Zbl

[33] **H Hofer**, **D A Salamon**, *Floer homology and Novikov rings*, from "The Floer memorial volume" (H Hofer, C H Taubes, A Weinstein, E Zehnder, editors), Progr. Math. 133, Birkhäuser, Basel (1995) 483–524 MR Zbl

[34] **T Honda**, *Formal groups and zeta-functions*, Osaka Math. J. 5 (1968) 199–213 MR Zbl

[35] **K Hori**, **S Katz**, **A Klemm**, **R Pandharipande**, **R Thomas**, **C Vafa**, **R Vakil**, **E Zaslow**, *Mirror symmetry*, Clay Mathematics Monographs 1, Amer. Math. Soc., Providence, RI (2003) MR Zbl

[36] **D Joyce**, *A generalization of manifolds with corners*, Adv. Math. 299 (2016) 760–862 MR Zbl

[37] **T Kimura**, **J Stasheff**, **A A Voronov**, *On operad structures of moduli spaces and string theory*, Comm. Math. Phys. 171 (1995) 1–25 MR Zbl

[38] **M Kontsevich**, **Y Manin**, *Gromov–Witten classes, quantum cohomology, and enumerative geometry*, Comm. Math. Phys. 164 (1994) 525–562 MR Zbl

[39] **M Lazard**, *Commutative formal groups*, Lecture Notes in Math. 443, Springer (1975) MR Zbl

[40] **Y Lekili**, **M Lipyanskiy**, *Geometric composition in quilted Floer theory*, Adv. Math. 236 (2013) 1–23 MR Zbl

[41] **Y Lekili**, **T Perutz**, *Arithmetic mirror symmetry for the 2–torus*, preprint (2012) arXiv 1211.4632

[42] **Y Lekili**, **A Polishchuk**, *Arithmetic mirror symmetry for genus 1 curves with n marked points*, Selecta Math. 23 (2017) 1851–1907 MR Zbl

[43] **The LMFDB collaboration**, *The L–functions and modular forms database*, electronic reference (2019) Available at `http://www.lfmdb.org`

[44] **V Lyubashenko**, *Category of A_∞–categories*, Homology Homotopy Appl. 5 (2003) 1–48 MR Zbl

[45] **V Lyubashenko**, **O Manzyuk**, *Unital A_∞–categories*, preprint (2008) arXiv 0802.2885

[46] **S MacLane**, *Homology*, 1st edition, Grundl. Math. Wissen. 114, Springer (1967) MR Zbl

[47] **S Ma'u**, *Quilted strips, graph associahedra, and A_∞ n–modules*, Algebr. Geom. Topol. 15 (2015) 783–799 MR Zbl

[48] **S Ma'u**, **K Wehrheim**, **C Woodward**, *A_∞ functors for Lagrangian correspondences*, Selecta Math. 24 (2018) 1913–2002 MR Zbl

[49] **S Ma'u**, **C Woodward**, *Geometric realizations of the multiplihedra*, Compos. Math. 146 (2010) 1002–1028 MR Zbl

[50] **J P May**, *A general algebraic approach to Steenrod operations*, from "The Steenrod algebra and its applications" (F P Peterson, editor), Lecture Notes in Math. 168, Springer (1970) 153–231 MR Zbl

[51] **J P May**, *The geometry of iterated loop spaces*, Lecture Notes in Math. 271, Springer (1972) MR Zbl

[52] **J H McClellan**, **T W Parks**, *Eigenvalue and eigenvector decomposition of the discrete Fourier transform*, IEEE Trans. Audio Electroacoust. AU-20 (1972) 66–74 MR

[53] **J E McClure**, **J H Smith**, *A solution of Deligne's Hochschild cohomology conjecture*, from "Recent progress in homotopy theory" (D M Davis, J Morava, G Nishida, W S Wilson, N Yagita, editors), Contemp. Math. 293, Amer. Math. Soc., Providence, RI (2002) 153–193 MR Zbl

[54] **D McDuff**, **D Salamon**, *J–holomorphic curves and quantum cohomology*, University Lecture Series 6, Amer. Math. Soc., Providence, RI (1994) MR Zbl

[55] **J J Millson**, *On the first Betti number of a constant negatively curved manifold*, Ann. of Math. 104 (1976) 235–247 MR Zbl

[56] **L D Olson**, *Hasse invariants and anomalous primes for elliptic curves with complex multiplication*, J. Number Theory 8 (1976) 397–414 MR Zbl

[57] **R Pandharipande**, *Rational curves on hypersurfaces (after A. Givental)*, from "Séminaire Bourbaki, Vol. 1997/98", Astérisque 252, Soc. Math. France, Paris (1998) Exposé 848, 307–340 MR Zbl

[58] **S Piunikhin**, **D Salamon**, **M Schwarz**, *Symplectic Floer–Donaldson theory and quantum cohomology*, from "Contact and symplectic geometry" (C B Thomas, editor), Publ. Newton Inst. 8, Cambridge Univ. Press (1996) 171–200 MR Zbl

[59] **M Reid**, *The complete intersection of two or more quadrics*, DPhil thesis, University of Cambridge (1972) Available at `https://homepages.warwick.ac.uk/~masda/3folds/qu.pdf`

[60] **A G Reznikov**, *Symplectic twistor spaces*, Ann. Global Anal. Geom. 11 (1993) 109–118 MR Zbl

[61] **Y Ruan**, **G Tian**, *Bott-type symplectic Floer cohomology and its multiplication structures*, Math. Res. Lett. 2 (1995) 203–219 MR Zbl

[62] **P Salvatore**, *Configuration spaces with summable labels*, from "Cohomological methods in homotopy theory" (J Aguadé, C Broto, C Casacuberta, editors), Progr. Math. 196, Birkhäuser, Basel (2001) 375–395 MR Zbl

[63] **M Schlessinger**, *Functors of Artin rings*, Trans. Amer. Math. Soc. 130 (1968) 208–222 MR Zbl

[64] **P Seidel**, *A_∞–subalgebras and natural transformations*, Homology Homotopy Appl. 10 (2008) 83–114 MR Zbl

[65] **P Seidel**, *Fukaya categories and Picard–Lefschetz theory*, Eur. Math. Soc., Zürich (2008) MR Zbl

[66] **P Seidel**, **N Wilkins**, *Covariant constancy of quantum Steenrod operations*, J. Fixed Point Theory Appl. 24 (2022) art. id. 52 MR Zbl

[67] **N Sheridan**, *On the Fukaya category of a Fano hypersurface in projective space*, Publ. Math. Inst. Hautes Études Sci. 124 (2016) 165–317 MR Zbl

[68] **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

[69] **I Smith**, *Floer cohomology and pencils of quadrics*, Invent. Math. 189 (2012) 149–250 MR Zbl

[70] **N E Steenrod**, *Cohomology operations*, Annals of Mathematics Studies 50, Princeton Univ. Press (1962) MR Zbl

[71] **V Tourtchine**, *Dyer–Lashof–Cohen operations in Hochschild cohomology*, Algebr. Geom. Topol. 6 (2006) 875–894 MR Zbl

[72] **A A Voronov**, **M Gerstenkhaber**, *Higher-order operations on the Hochschild complex*, Funktsional. Anal. i Prilozhen. 29 (1995) 1–6 MR Zbl In Russian; translated in Funct. Anal. Appl. 29 (1995) 1–5

[73] **C A Weibel**, *An introduction to homological algebra*, Cambridge Studies in Advanced Mathematics 38, Cambridge Univ. Press (1994) MR Zbl

[74] **N Wilkins**, *A construction of the quantum Steenrod squares and their algebraic relations*, Geom. Topol. 24 (2020) 885–970 MR Zbl

[75] **N Wilkins**, *Quantum Steenrod squares and the equivariant pair-of-pants in symplectic cohomology*, J. Topol. Anal. 15 (2023) 1–56 MR Zbl

[76] **J Young**, *Brace bar–cobar duality*, preprint (2013) arXiv 1309.2820

[77] **A Zimmermann**, *Fine Hochschild invariants of derived categories for symmetric algebras*, J. Algebra 308 (2007) 350–367 MR Zbl

Department of Mathematics, Massachusetts Insitutute of Technology
Cambridge, MA, United States

pseidel@mit.edu

Proposed: Ciprian Manolescu Received: 21 October 2019
Seconded: Leonid Polterovich, Yakov Eliashberg Revised: 14 January 2022

Geometry & Topology **27**:8 (2023) 3061–3094
DOI: 10.2140/gt.2023.27.3061
Published: 9 November 2023

AGT relations for sheaves on surfaces

ANDREI NEGUȚ

We consider a natural generalization of the Carlsson–Okounkov Ext operator on the K–theory groups of the moduli spaces of stable sheaves on a smooth projective surface. We compute the commutation relations between the Ext operator and the action of the deformed W–algebra on K–theory, which was developed by the author in previous work. The conclusion is that the Ext operator is closely related to a vertex operator, thus giving a mathematical incarnation of the Alday–Gaiotto–Tachikawa correspondence for a general algebraic surface.

14J60; 14D21

1 Introduction

1.1 Fix a smooth projective surface S over an algebraically closed field of characteristic zero (henceforth denoted by $\mathbb{C}$), and invariants $(r, c_1) \in \mathbb{N} \times H^2(S, \mathbb{Z})$. An important object in algebraic geometry is the moduli space

$$\mathcal{M} = \bigsqcup_{c_2=\lceil((r-1)/2r)c_1^2\rceil}^{\infty} \mathcal{M}_{c_2} \tag{1-1}$$

of H–stable sheaves on S with invariants (r, c_1, c_2) for any $c_2 \in \mathbb{Z}$. The reason that c_2 is bounded below is called Bogomolov's inequality, which states that there are no H–stable sheaves if $c_2 < ((r-1)/2r)c_1^2$. We make the same assumptions as in our earlier work [15; 17; 16]:

- **Assumption A** $\gcd(r, c_1 \cdot H) = 1$.
- **Assumption S** Either $\omega_S \cong \mathcal{O}_S$, or $c_1(\omega_S) \cdot H < 0$.

Assumption A implies that $\mathcal{M}$ is proper and there exists a universal sheaf[1]

$$\begin{tikzcd} \mathcal{U} \arrow[d, dotted] \\ \mathcal{M} \times S \end{tikzcd} \tag{1-2}$$

Assumption S implies that $\mathcal{M}$ is smooth.

1.2 The enumerative geometry of the moduli space of stable sheaves is quite rich, as evidenced by Donaldson invariants arising as certain integrals of cohomology classes on $\mathcal{M}$. In the present paper, we will consider algebraic K–theory instead of cohomology, a process which accounts for the adjective "deformed" in the representation-theoretic structures explained in Section 1.6. Explicitly, we consider the following algebraic K–theory groups with $\mathbb{Q}$ coefficients:

$$K_{\mathcal{M}} = \bigoplus_{c_2 = \lceil ((r-1)/2r) c_1^2 \rceil}^{\infty} K_0(\mathcal{M}_{c_2}) \otimes_{\mathbb{Z}} \mathbb{Q}. \tag{1-3}$$

Let $m \in \mathrm{Pic}(S)$, and consider two copies $\mathcal{M}$ and $\mathcal{M}'$ of the moduli space (1-1). These two copies may be defined with respect to a different c_1 and stability condition H, but we assume that the rank r of the sheaves parametrized by $\mathcal{M}$ and $\mathcal{M}'$ is the same. In this paper, we will mostly be concerned with the virtual vector bundle

$$\begin{tikzcd} & \mathcal{E}_m \arrow[d, dashed] & \\ & \mathcal{M} \times \mathcal{M}' \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_2"] & \\ \mathcal{M} & & \mathcal{M}' \end{tikzcd} \tag{1-4}$$

(a straightforward generalization of the construction of Carlsson and Okounkov [7]) given by

$$\mathcal{E}_m = \mathrm{R}\Gamma(m) - \mathrm{R}\pi_*(\mathrm{R}\mathscr{H}om(\mathcal{U}', \mathcal{U} \otimes m)). \tag{1-5}$$

The $\mathrm{R}\mathscr{H}om$ is computed on $\mathcal{M} \times \mathcal{M}' \times S$: the notation $\mathcal{U}$, $\mathcal{U}'$ and m stands for the pullback of the universal sheaf from $\mathcal{M} \times S$ and $\mathcal{M}' \times S$, respectively, as well as the pullback of the line bundle m from S. Similarly, $\pi : \mathcal{M} \times \mathcal{M}' \times S \to \mathcal{M} \times \mathcal{M}'$ is the standard projection, so $\mathcal{E}_m$ is a complex of coherent sheaves on $\mathcal{M} \times \mathcal{M}'$.

[1] We require the universal sheaves on the various connected components of $\mathcal{M}$ to be constructed as in [15, Section 5.9], which will ensure that they lift in a compatible way to the moduli spaces $\mathfrak{Z}_1$ and $\mathfrak{Z}_2^\bullet$ of Section 2.4.

1.3 Any Schur functor applied to $\mathcal{E}_m$ gives rise to a K–theory class on $\mathcal{M} \times \mathcal{M}'$, which in turn induces an operator from $K_{\mathcal{M}'}$ to $K_{\mathcal{M}}$ via the usual formalism of correspondences. With this in mind, let us consider the following immediate generalization of Carlsson and Okounkov [7, Equation (3)] and Carlsson, Nekrasov and Okounkov [6, Equation (19)].

Definition 1.4 Consider the so-called *Ext operator* $K_{\mathcal{M}'} \xrightarrow{A_m} K_{\mathcal{M}}$ given by

$$A_m = \pi_{1*}(\wedge^\bullet \mathcal{E}_m \cdot \pi_2^*), \tag{1-6}$$

with π_1 and π_2 as in (1-4). The pushforward and pullback maps are well-defined due to the properness and smoothness of $\mathcal{M}$ and $\mathcal{M}'$, respectively.

In (1-6), the symbol $\wedge^\bullet \mathcal{E}_m$ denotes the total exterior power of $\mathcal{E}_m$; as $\mathcal{E}_m$ is in general a complex of coherent sheaves, some explanation is in order. Specifically, consider

$$\wedge^\bullet\left(\frac{\mathcal{E}_m}{t}\right) = \sum_{k \geq 0} (-t)^{-k} [\wedge^k \mathcal{E}_m] \in K_{\mathcal{M}\times\mathcal{M}'}[\![t^{-1}]\!], \tag{1-7}$$

where the right-hand side is the power series expansion of a rational function in t; see Section 3.1 for details. Then the quantity $\wedge^\bullet \mathcal{E}_m$ in (1-6) denotes the $t = 1$ specialization of (1-7). If this specialization is not well-defined, then all the results in Sections 1.6 and 1.9 hold with m replaced by m/t, and with all formulas being equalities of rational functions in t; see Section 3.1 for details.

Example 1.5 Let $\mathcal{M} = \mathcal{M}'$ and $m = \mathcal{O}_S/t$, with t being a formal parameter. Then Assumption S implies that $\mathcal{E}_{\mathcal{O}_S}$ is locally free (up to a constant sheaf) and that

$$\mathcal{E}_{\mathcal{O}_S}|_\Delta \cong \mathrm{Tan}_{\mathcal{M}},$$

where $\Delta \subset \mathcal{M} \times \mathcal{M}'$ denotes the diagonal. By a simple computation involving correspondences, the isomorphism above implies that

$$\mathrm{Tr}(A_{\mathcal{O}_S/t}) = \sum_{k \geq 0} (-t)^{-k} \chi(\mathcal{M}, \wedge^k \mathrm{Tan}_{\mathcal{M}})$$

(up to a constant rational function in t). The right-hand side is the χ_t–genus of the moduli space $\mathcal{M}$, as considered for example in Göttsche and Kool [10].

1.6 In the present paper, we will seek to determine the Ext operator A_m using the representation-theoretic properties of the vector space $K_{\mathcal{M}}$. To this end, we need to make $K_{\mathcal{M}}$ into a representation of an appropriate algebra which is "big" enough, in order to constrain the operator A_m as much as possible. A candidate for such an algebra is $\mathcal{A}_r$, namely a particular integral form of the deformed W–algebra of type $\mathfrak{gl}_r$ (initially defined in Awata, Kubo, Odake and Shiraishi [1] and Feigin and Frenkel [8]). The main purpose of our work in [15; 17; 16] is to construct an action $\mathcal{A}_r \curvearrowright K_{\mathcal{M}}$; we will recall the construction in Section 2, but let us summarize the main idea here. In [17, Section 6.7], we construct certain geometric operators

$$K_{\mathcal{M}} \xrightarrow{W_{n,k}} K_{\mathcal{M}\times S} \quad \text{for all } (n,k) \in \mathbb{Z} \times \mathbb{N}. \tag{1-8}$$

Under Assumptions A and S, we show in [16, Theorem 4.15] that the operators $W_{n,k}$ satisfy the quadratic commutation relations developed in [1] and [8]; see (2-28) for the specific form of these relations in our language. In [17, Theorem 6.9], we further show that $W_{n,k} = 0$ for all $n \in \mathbb{Z}$ and $k > r$, which tautologically implies that the operators (1-8) yield an action $\mathcal{A}_r \curvearrowright K_{\mathcal{M}}$. Write

$$q = [\omega_S] \in K_S := K_0(S) \otimes_{\mathbb{Z}} \mathbb{Q}. \tag{1-9}$$

Given two copies $\mathcal{M}$ and $\mathcal{M}'$ of the moduli space of stable sheaves, each with its own universal sheaf $\mathcal{U}$ and $\mathcal{U}'$, respectively, we may write

$$u = \det \mathcal{U} \quad \text{and} \quad u' = \det \mathcal{U}' \tag{1-10}$$

for the determinant line bundles on $\mathcal{M} \times S$ and $\mathcal{M}' \times S$, respectively. We set

$$\gamma = \frac{m^r u}{q^r u'}, \tag{1-11}$$

which is the class of a line bundle on $\mathcal{M} \times \mathcal{M}' \times S$ (it is implicit that m and q are pulled back from S). Our main result, which will be proved in Section 3, is:

Theorem 1.7 *We have the following interaction between the Ext operator* (1-6) *and the generators* (1-8) *of the W–algebra action*:

$$A_m W_k(x)(1-x) = m^k W_k(x\gamma) A_m \left(1 - \frac{x}{q^k}\right), \tag{1-12}$$

where $W_k(x) = \sum_{n \in \mathbb{Z}} W_{n,k}/x^n$. The series coefficients of the two sides of (1-12) *are maps $K_{\mathcal{M}'} \to K_{\mathcal{M}\times S}$ which arise from certain correspondences in $K_{\mathcal{M}\times\mathcal{M}'\times S}$.*

Remark See Section 2.1 for a review of correspondences as K–theoretic operators. In particular, the composition of operators depends on which of A_m and $W_k(x)$ is on the left of the other:

$$A_m W_{n,k} \colon K_{\mathcal{M}'} \xrightarrow{W_{n,k}} K_{\mathcal{M}' \times S} \xrightarrow{A_m \times \mathrm{Id}_S} K_{\mathcal{M} \times S},$$
$$W_{n,k} A_m \colon K_{\mathcal{M}'} \xrightarrow{A_m} K_{\mathcal{M}} \xrightarrow{W_{n,k}} K_{\mathcal{M} \times S}.$$

The expressions above are actually given by certain correspondences in $K_{\mathcal{M} \times \mathcal{M}' \times S}$. Then the factors q and γ on the right-hand side of (1-12) indicate multiplication of the aforementioned correspondences by various powers of the line bundles (1-9) and (1-11).

1.8 A major motivation for the study of the Ext operator A_m stems from mathematical physics: as explained in Carlsson, Nekrasov and Okounkov [6], the operator A_m encodes the contribution of bifundamental matter to partition functions of 5d supersymmetric gauge theory on the algebraic surface S times a circle. Moreover, the deformed W–algebra $\mathcal{A}_r$ encodes symmetries of Toda conformal field theory. In this language, (1-12) becomes a mathematical manifestation of the Alday–Gaiotto–Tachikawa (AGT) correspondence between gauge theory and conformal field theory, by describing the Ext operator A_m in terms of its commutation with W–algebra generators. To the author's knowledge, the present paper is the first mathematical treatment of AGT over general algebraic surfaces in rank $r > 1$ (the reference [6] used different techniques from ours to describe the Ext operator in the $r = 1$ case).

However, we note that formulas (1-12) are not enough to completely determine A_m for a general smooth projective surface S, and one should instead work with a deformed vertex operator algebra which properly contains several deformed W–algebras $\mathcal{A}_r$. In the nondeformed case, a potential candidate for such a larger algebra was studied in Feigin and Gukov [9], where the authors expect that it contains operators which modify sheaves on S along entire curves, on top of our operators $W_{n,k}$ which modify sheaves at individual points. While we give a complete algebrogeometric description of the latter operators, we do not have such a description for the former operators. Once such a description is available, we hope that one can extend Theorem 1.7 to a bigger vertex operator algebra properly containing $\mathcal{A}_r$.

There is a situation where formulas (1-12) do indeed determine the Ext operator A_m completely: this corresponds to taking $S = \mathbb{A}^2$, replacing $\mathcal{M}$ by the moduli space of framed rank r sheaves on the projective plane, and working with torus equivariant K–theory; see Section 4.1 for details. In this particular case, we showed in [14] that $K_{\mathcal{M}}$ is isomorphic to the universal Verma module of $\mathcal{A}_r$. Theorem 1.7 holds in the situation at hand, and we will show in Theorem 4.5 that our formulas completely determine the Ext operator A_m. This precisely yields the AGT correspondence for 5d supersymmetric gauge theory on $\mathbb{A}^2 \times S^1$; see for instance Braverman, Finkelberg and Nakajima [4], Bruzzo, Pedrini, Sala and Szabo [5], Maulik and Okounkov [12] and Schiffmann and Vasserot [18] for the history of this correspondence in mathematical language.

1.9 Alongside the operators (1-8), we constructed in [16, Theorem 4.15] K–theory lifts of the operators introduced by Grojnowski and Nakajima [11; 13] for $r = 1$, and generalized by Baranovsky [2] for any r, in cohomology:

$$K_{\mathcal{M}} \xrightarrow{P_n} K_{\mathcal{M}\times S} \quad \text{for all } n \in \mathbb{Z}\backslash 0. \tag{1-13}$$

These operators satisfy the Heisenberg commutation relation (2-29), and interact with the deformed W–algebra generators according to relation (2-30).

Recall the line bundles q and γ of (1-9) and (1-11), respectively, and the footnote in Theorem 1.7 to properly interpret compositions of the operators A_m and $P_{\pm n}$.

Theorem 1.10 *We have the following interaction between the Ext operator* (1-6) *and the Heisenberg operators $P_{\pm n}$ for all $n > 0$:*

$$A_m P_{-n} - P_{-n} A_m \gamma^n = A_m(1-\gamma^n), \tag{1-14}$$

$$A_m P_n - P_n A_m \gamma^{-n} = A_m(\gamma^{-n} - q^{rn}). \tag{1-15}$$

In $\mathcal{A}_r$, the series $W_r(x)$ matches the normal-ordered exponential of the generating series of the P_n; see Theorem 2.8. With this in mind, it is straightforward to show that the $k = r$ case of Theorem 1.7 follows from Theorem 1.10.

For any $\alpha \in K_S$, we will write $P_n\{\alpha\}$ for the composition

$$P_n\{\alpha\}\colon K_{\mathcal{M}} \xrightarrow{P_n} K_{\mathcal{M}\times S} \xrightarrow{\text{multiplication by } \mathrm{proj}_2^*(\alpha)} K_{\mathcal{M}\times S} \xrightarrow{\mathrm{proj}_{1*}} K_{\mathcal{M}},$$

where proj_1 and proj_2 are the projections from $\mathcal{M}\times S$ to $\mathcal{M}$ and S, respectively. Let q_1 and q_2 denote the Chern roots of the cotangent bundle Ω_S^1. Any symmetric Laurent polynomial in q_1 and q_2 gives rise to a well-defined element of K_S, via

$$q_1 + q_2 = [\Omega_S^1] \quad \text{and} \quad q = q_1 q_2 = [\omega_S].$$

Define

$$\Phi_m = A_m \exp\left[\sum_{n=1}^{\infty} \frac{P_n}{n}\left\{\frac{(q^n-1)q^{-nr}}{[n]_{q_1}[n]_{q_2}}\right\}\right], \tag{1-16}$$

where $[n]_x = 1 + x + \cdots + x^{n-1}$. The expression in curly brackets is an element of K_S because $[n]_{q_1}[n]_{q_2}$ is a unit in the ring K_S.

Remark To see that $[n]_{q_1}[n]_{q_2}$ is a unit in the ring K_S, since the Chern character gives us an isomorphism $K_S \cong A^*(S,\mathbb{Q})$, we have $q_1+q_2=[\Omega^1_S]\in 2+\mathcal{N}$ and $q=[\omega_S]\in 1+\mathcal{N}$, where $\mathcal{N}\subset K_S$ denotes the nilradical. Therefore $[n]_{q_1}[n]_{q_2}\in n^2+\mathcal{N}$, and is thus invertible in the ring K_S.

Corollary 1.11 *Formulas* (1-12), (1-14) *and* (1-15) *imply*

$$[\Phi_m W_k(x) - m^k W_k(x\gamma)\Phi_m]\left(1-\frac{x}{q^k}\right)=0, \tag{1-17}$$

$$\Phi_m P_{\pm n} - P_{\pm n}\Phi_m\gamma^{\mp n} = \pm\Phi_m(\gamma^{\mp n}-q^{\pm rn}) \tag{1-18}$$

for all $k,n>0$. An operator Φ_m satisfying (1-17) *and* (1-18) *is called a* ***vertex operator****.*

Acknowledgements I thank Boris Feigin, Sergei Gukov, Hiraku Nakajima, Nikita Nekrasov, Andrei Okounkov, Francesco Sala and Alexander Tsymbaliuk for many interesting discussions on the subject of Ext operators and W–algebras. I gratefully acknowledge the support of NSF grant DMS–1600375.

2 The moduli space of sheaves

2.1 Throughout the present paper, we will work with smooth projective varieties over the field $\mathbb{C}$. For such varieties X, we let

$$K_X = K_0(X)\otimes_{\mathbb{Z}}\mathbb{Q}$$

be the Grothendieck group of the category of coherent sheaves on X, with scalars extended to $\mathbb{Q}$. Derived tensor product yields a ring structure on K_X, and we have pullback and pushforward maps for any proper l.c.i. morphism $X\to Y$.

Definition 2.2 Given smooth projective varieties X and Y, any class $\Gamma\in K_{X\times Y}$ (called a "correspondence" in this setup) defines an operator

$$K_Y \xrightarrow{\Psi_\Gamma} K_X,\quad \Psi_\Gamma = \mathrm{proj}_{X*}(\Gamma\cdot\mathrm{proj}_Y^*), \tag{2-1}$$

where $\mathrm{proj}_X,\mathrm{proj}_Y$ denote the projection maps from $X\times Y$ to X and Y, respectively.

The composition of operators (2-1) can also be described as a correspondence

$$\Psi_\Gamma\circ\Psi_{\Gamma'} = \Psi_{\Gamma''}\colon K_Z\to K_X \tag{2-2}$$

for any $\Gamma \in K_{X\times Y}$ and $\Gamma' \in K_{Y\times Z}$, where

$$\Gamma'' = \mathrm{proj}_{X\times Z*}\big(\mathrm{proj}^*_{X\times Y}(\Gamma) \otimes \mathrm{proj}^*_{Y\times Z}(\Gamma')\big), \tag{2-3}$$

where $\mathrm{proj}_{X\times Y}$, $\mathrm{proj}_{Y\times Z}$ and $\mathrm{proj}_{X\times Z}$ are the standard projections from $X\times Y\times Z$ to $X\times Y$, $Y\times Z$ and $X\times Z$. Throughout the present paper, all operators $K_Y \to K_X$ arise from explicit correspondences. While we will use the language of composition of operators for convenience, what is really happening behind the scenes is composition of correspondences under the operation $(\Gamma,\Gamma') \mapsto \Gamma''$ of (2-3).

2.3 In Section 1.6, we referred to various operators $K_{\mathcal{M}} \to K_{\mathcal{M}\times S}$ as defining an action of a certain algebra on $K_{\mathcal{M}}$, and we will now explain the meaning of this notion. Given two arbitrary homomorphisms (of abelian groups)

$$K_{\mathcal{M}} \xrightarrow{x,y} K_{\mathcal{M}\times S}, \tag{2-4}$$

their "product" $xy|_\Delta$ is defined as the composition

$$xy|_\Delta \colon K_{\mathcal{M}} \xrightarrow{y} K_{\mathcal{M}\times S} \xrightarrow{x\times \mathrm{Id}_S} K_{\mathcal{M}\times S\times S} \xrightarrow{\mathrm{Id}_{\mathcal{M}}\times\Delta^*} K_{\mathcal{M}\times S}$$

where $S \xrightarrow{\Delta} S\times S$ is the diagonal. It is easy to check that $(xy|_\Delta)z|_\Delta = x(yz|_\Delta)|_\Delta$, hence the aforementioned notion of product is associative, and it makes sense to define $x_1\cdots x_n|_\Delta$ for arbitrarily many operators $x_1,\dots,x_n\colon K_{\mathcal{M}} \to K_{\mathcal{M}\times S}$.

Similarly, given two operators (2-4), we may define their commutator

$$K_{\mathcal{M}} \xrightarrow{[x,y]} K_{\mathcal{M}\times S\times S}$$

as the difference of the two compositions

$$K_{\mathcal{M}} \xrightarrow{y} K_{\mathcal{M}\times S} \xrightarrow{x\times\mathrm{Id}_S} K_{\mathcal{M}\times S\times S},$$
$$K_{\mathcal{M}} \xrightarrow{x} K_{\mathcal{M}\times S} \xrightarrow{y\times\mathrm{Id}_S} K_{\mathcal{M}\times S\times S} \xrightarrow{\mathrm{Id}_{\mathcal{M}}\times\mathrm{swap}^*} K_{\mathcal{M}\times S\times S},$$

where $\mathrm{swap}\colon S\times S \to S\times S$ is the permutation of the two factors. In all cases studied in the present paper, we will have[2]

$$[x,y] = \Delta_*(z)$$

for some $K_{\mathcal{M}} \xrightarrow{z} K_{\mathcal{M}\times S}$ which is uniquely determined (the diagonal embedding Δ_* is injective because it has a left inverse), and which will be denoted by $z = [x,y]_{\mathrm{red}}$. We leave it as an exercise to the interested reader to prove that the commutator satisfies

[2]Here we abuse notation by writing Δ_* instead of $(\mathrm{Id}_{\mathcal{M}}\times\Delta)_*$ for the diagonal map $K_{\mathcal{M}\times S} \to K_{\mathcal{M}\times S\times S}$.

the Leibniz rule in the form $[xy|_\Delta, z]_{\mathrm{red}} = x[y,z]_{\mathrm{red}}|_\Delta + [x,z]_{\mathrm{red}}y|_\Delta$, and the Jacobi identity in the form $\big[[x,y]_{\mathrm{red}}, z\big]_{\mathrm{red}} + \big[[y,z]_{\mathrm{red}}, x\big]_{\mathrm{red}} + \big[[z,x]_{\mathrm{red}}, y\big]_{\mathrm{red}} = 0$.

Finally, we consider the ring homomorphism $\mathbb{K} = \mathbb{Z}[q_1^{\pm 1}, q_2^{\pm 1}]^{\mathrm{Sym}} \to K_S$ given by sending q_1 and q_2 to the Chern roots of the cotangent bundle of S (therefore, $q = q_1q_2$ goes to the class of the canonical line bundle). We will often abuse notation, and write q_1, q_2, q for the images of the indeterminates in the ring K_S. For any $\lambda \in \mathbb{K}$ and any operator (2-4), we may define their product as the composition

$$\lambda \cdot x \colon K_{\mathcal{M}} \xrightarrow{x} K_{\mathcal{M}\times S} \xrightarrow{\mathrm{Id}_{\mathcal{M}} \times (\text{multiplication by } \lambda)} K_{\mathcal{M}\times S},$$

where we identify $\lambda \in \mathbb{K}$ with its image in K_S. With this in mind, the ring K_S can be thought of as the "ring of constants" for the algebra of operators (2-4).

2.4 Recall the universal sheaf (1-2), and consider the derived scheme

$$\mathfrak{Z}_1 = \mathbb{P}_{\mathcal{M}\times S}(\mathcal{U}) \to \mathcal{M} \times S. \tag{2-5}$$

Since $\mathcal{U}$ is isomorphic to a quotient $\mathcal{V}/\mathcal{W}$ of vector bundles on $\mathcal{M} \times S$ (Proposition 2.2 of [15]), the projectivization in (2-5) is defined as the derived zero locus of a section of a vector bundle on the projective bundle $\mathbb{P}_{\mathcal{M}\times S}(\mathcal{V})$. However, it was shown in [15, Proposition 2.10] that under Assumption S, the derived zero locus is actually a smooth scheme

$$\mathfrak{Z}_1 = \bigsqcup_{c=\lceil((r-1)/2r)c_1^2\rceil}^{\infty} \mathfrak{Z}_{c+1,c},$$

whose connected components are given by

$$\mathfrak{Z}_{c+1,c} = \{(\mathcal{F}_{c+1}, \mathcal{F}_c) \text{ such that } \mathcal{F}_{c+1} \subset_x \mathcal{F}_c \text{ for some } x \in S\} \subset \mathcal{M}_{c+1} \times \mathcal{M}_c, \tag{2-6}$$

and $\mathcal{F}' \subset_x \mathcal{F}$ means that $\mathcal{F}' \subset \mathcal{F}$ and the quotient $\mathcal{F}/\mathcal{F}'$ is isomorphic to the length one skyscraper sheaf at the point $x \in S$. This scheme comes with projection maps

$$\begin{tikzcd} & \mathfrak{Z}_{c+1,c} \arrow[dl, "p_+"'] \arrow[d, "p_S"] \arrow[dr, "p_-"] & \\ \mathcal{M}_{c+1} & S & \mathcal{M}_c \end{tikzcd} \tag{2-7}$$

More generally, we defined a derived scheme $\mathfrak{Z}_2^\bullet$ in [17, Definition 4.17], which was shown (under Assumption S, in [17, Proposition 4.21]) to be a smooth scheme

$$\mathfrak{Z}_2^\bullet = \bigsqcup_{c=\lceil((r-1)/2r)c_1^2\rceil}^{\infty} \mathfrak{Z}_{c+2,c}^\bullet,$$

whose connected components are given by

(2-8) $\mathfrak{Z}^\bullet_{c+2,c} = \{(\mathcal{F}_{c+2} \subset_x \mathcal{F}_{c+1} \subset_x \mathcal{F}_c) \text{ for some } x \in S\} \subset \mathcal{M}_{c+2} \times \mathcal{M}_{c+1} \times \mathcal{M}_c$.

This scheme is equipped with projection maps as in (2-9) below, but we observe that the rhombus is not derived Cartesian (and this is key to our construction):

$$\begin{tikzcd}[column sep=small] & \mathfrak{Z}^\bullet_{c+2,c} \arrow[dl, "\pi_+"'] \arrow[dr, "\pi_-"] & \\ \mathfrak{Z}_{c+2,c+1} \arrow[dr, "p_-\times p_S"'] & & \mathfrak{Z}_{c+1,c} \arrow[dl, "p_+\times p_S"] \\ & \mathcal{M}_{c+1} \times S & \end{tikzcd} \tag{2-9}$$

Note that all of the maps in the diagram above are proper, l.c.i. morphisms. Define

$$\mathfrak{Z}^\bullet_n = \bigsqcup_{c=\lceil ((r-1)/2r)c_1^2 \rceil}^{\infty} \mathfrak{Z}^\bullet_{c+n,c}, \tag{2-10}$$

whose connected components are given by derived fiber products

(2-11) $\mathfrak{Z}^\bullet_{c+n,c} = \mathfrak{Z}^\bullet_{c+n,c+n-2} \underset{\mathfrak{Z}_{c+n-1,c+n-2}}{\times} \cdots \underset{\mathfrak{Z}_{c+2,c+1}}{\times} \mathfrak{Z}^\bullet_{c+2,c} \to \mathcal{M}_{c+n} \times \cdots \times \mathcal{M}_c$.

While $\mathfrak{Z}^\bullet_n$ is a derived scheme, we note that its closed points are all of the form

(2-12) $\mathfrak{Z}^\bullet_{c+n,c} = \{(\mathcal{F}_{c+n}, \dots, \mathcal{F}_c) \text{ sheaves with } \mathcal{F}_{c+n} \subset_x \cdots \subset_x \mathcal{F}_c \text{ for some } x \in S\}$.

Therefore, we have the following projection maps, which only remember the smallest and the largest sheaf in a flag (2-12):

$$\begin{tikzcd} & \mathfrak{Z}^\bullet_{c+n,c} \arrow[dl, "p_+"'] \arrow[d, "p_S"] \arrow[dr, "p_-"] & \\ \mathcal{M}_{c+n} & S & \mathcal{M}_c \end{tikzcd} \tag{2-13}$$

(the notation generalizes (2-7)). In diagram (2-13), the maps $p_\pm$ are l.c.i. morphisms, and the maps $p_\pm \times p_S$ are proper (they inherit these properties from the maps in (2-9)). Finally, we consider the line bundles $\mathcal{L}_1, \dots, \mathcal{L}_n$ on $\mathfrak{Z}^\bullet_n$, whose fibers are given by

$$\mathcal{L}_i|_{(\mathcal{F}_{c+n},\dots,\mathcal{F}_c)} = \mathcal{F}_{c+n-i,x}/\mathcal{F}_{c+n-i+1,x} \tag{2-14}$$

on the connected component $\mathfrak{Z}^\bullet_{c+n,c} \subset \mathfrak{Z}^\bullet_n$.

2.5 Using the derived scheme (2-11) and the maps (2-13), define for all $n, k \in \mathbb{N}$

$$K_{\mathcal{M}} \xrightarrow{L_{n,k}} K_{\mathcal{M}\times S}, \qquad L_{n,k} = (-1)^{k-1}(p_+ \times p_S)_*(\mathcal{L}_n^k \cdot p_-^*), \tag{2-15}$$

$$K_{\mathcal{M}} \xrightarrow{U_{n,k}} K_{\mathcal{M}\times S}, \qquad U_{n,k} = \frac{(-1)^{rn+k-1}u^n}{q^{(r-1)n}}(p_- \times p_S)_*\left(\frac{\mathcal{L}_n^k}{\mathcal{Q}^r} \cdot p_+^*\right), \tag{2-16}$$

where $\mathcal{Q} = \mathcal{L}_1 \cdots \mathcal{L}_n$, and u is the determinant of the universal sheaf on $\mathcal{M} \times S$, as in (1-10).[3] Implicit in the definitions (2-15) and (2-16) is that we define the operators therein for all components $\mathcal{M}_c$ of the moduli space $\mathcal{M}$. We also set

$$L_{n,0} = U_{n,0} = \delta_n^0 \quad \text{and} \quad L_{0,k} = U_{0,k} = \delta_k^0. \tag{2-17}$$

Finally, consider for all $k \in \mathbb{N} \sqcup 0$ the operators

$$E_k \colon K_{\mathcal{M}} \xrightarrow{\text{pullback}} K_{\mathcal{M}\times S} \xrightarrow{\text{multiplication by } \wedge^k \mathcal{U}} K_{\mathcal{M}\times S}. \tag{2-18}$$

Since $\mathcal{U} \cong \mathcal{V}/\mathcal{W}$ is a coherent sheaf of projective dimension one on $\mathcal{M} \times S$ (see [15, Proposition 2.2]), the class $\wedge^k \mathcal{U}$ in (2-18) is defined by setting

$$\wedge^\bullet\left(\frac{\mathcal{U}}{z}\right) = \frac{\wedge^\bullet\left(\frac{\mathcal{V}}{z}\right)}{\wedge^\bullet\left(\frac{\mathcal{W}}{z}\right)} \tag{2-19}$$

and picking out the coefficient of z^{-k} when expanding in negative powers of z. The reason for our notation for the operators (2-15), (2-16) and (2-18) is that these three operators are respectively lower triangular, upper triangular, and diagonal with respect to the grading on $K_{\mathcal{M}}$ by the second Chern class; see (1-3).

Definition 2.6 [17, Section 6.7] For any $(n, k) \in \mathbb{Z} \times \mathbb{N}$, consider the operators

$$W_{n,k} = \sum_{k_0+k_1+k_2=k}^{n_2-n_1=n} q^{(k-1)n_2} \cdot L_{n_1,k_1} E_{k_0} U_{n_2,k_2}\Big|_\Delta \tag{2-20}$$

as k_0, k_1, k_2, n_1, n_2 run over $\mathbb{N} \sqcup 0$ (recall the convention (2-17)).

Note that (2-20) is an infinite sum, but its action on $K_{\mathcal{M}}$ is well-defined because the operators $L_{n,k}$ (resp. $U_{n,k}$) increase (resp. decrease) the c_2 of stable sheaves by n, and Bogomolov's inequality ensures that the moduli space of stable sheaves is empty if c_2 is small enough.

[3] Note that u parametrizes the determinant of any one of the sheaves $\mathcal{F}_{c+n}, \dots, \mathcal{F}_c$ in a flag (2-12), since these sheaves have canonically isomorphic determinants; see Proposition 3.4.

2.7 Similarly with (2-15) and (2-16), for all $n \in \mathbb{N}$ we have the operators

$$K_{\mathcal{M}} \xrightarrow{P_{-n}} K_{\mathcal{M}\times S}, \quad P_{-n} = (p_+ \times p_S)_*\biggl(\sum_{i=0}^{n-1} \frac{q^i \mathcal{L}_n}{\mathcal{L}_{n-i}} \cdot p_-^*\biggr), \tag{2-21}$$

$$K_{\mathcal{M}} \xrightarrow{H_{-n}} K_{\mathcal{M}\times S}, \quad H_{-n} = (p_+ \times p_S)_*(p_-^*), \tag{2-22}$$

$$K_{\mathcal{M}} \xrightarrow{P_{n}} K_{\mathcal{M}\times S}, \quad P_{n} = (-1)^{rn} u^n (p_- \times p_S)_*\biggl(\sum_{i=0}^{n-1} \frac{q^i \mathcal{L}_n}{\mathcal{Q}^r \mathcal{L}_{n-i}} \cdot p_+^*\biggr), \tag{2-23}$$

$$K_{\mathcal{M}} \xrightarrow{H_{n}} K_{\mathcal{M}\times S}, \quad H_{n} = (-1)^{rn} u^n (p_- \times p_S)_*(\mathcal{Q}^{-r} \cdot p_+^*). \tag{2-24}$$

As a consequence of [17, formulas (2.15) and (5.18)–(5.21)], the operators $H_{\pm n}$ are to the operators $P_{\pm n}$ as complete symmetric functions are to power sum functions

$$\sum_{n=0}^{\infty} \frac{H_{\pm n}}{z^{\pm n}} = \exp\biggl(\sum_{n=1}^{\infty} \frac{P_{\pm n}}{n z^{\pm n}}\biggr)\bigg|_{\Delta} \tag{2-25}$$

or, explicitly,

$$H_0 = \mathrm{proj}_1^*,$$

where $\mathrm{proj}_1 \colon \mathcal{M} \times S \to \mathcal{M}$ is the usual projection, and

$$\begin{aligned} H_{\pm 1} &= P_{\pm 1}, \\ H_{\pm 2} &= \tfrac{1}{2}(P_{\pm 1} P_{\pm 1}|_{\Delta} + P_{\pm 2}), \\ H_{\pm 3} &= \tfrac{1}{6}(P_{\pm 1} P_{\pm 1} P_{\pm 1}|_{\Delta} + 3 P_{\pm 1} P_{\pm 2}|_{\Delta} + 2 P_{\pm 3}), \end{aligned}$$

and so on.

Theorem 2.8 [17, Theorem 6.9] *The operators* (2-20) *satisfy*

$$W_{n,r} = u \sum_{n_1,n_2 \geq 0}^{n_2 - n_1 = n} H_{-n_1} H_{n_2}\bigg|_{\Delta} \quad \textit{for all } n \in \mathbb{Z}, \tag{2-26}$$

$$W_{n,k} = 0 \quad \textit{for all } k > r. \tag{2-27}$$

2.9 We will now present the interaction of the operators (2-20), (2-21) and (2-23). Recall the commutator construction from Section 2.3.

The following theorem was stated in [17, Theorem 3.13 and Proposition 3.15] and proved in [16, Theorem 4.15] under Assumption S.

Theorem 2.10 *We have the following formulas for all $n, n' \in \mathbb{Z}$ and $k, k' \in \mathbb{N}$:*

$$[W_{n,k}, W_{n',k'}] = \Delta_*\left(\sum_{\substack{m/l \le m'/l' \\ \min(l,l') \le \min(k,k') \\ k+k'=l+l' \\ m+m'=n+n'}} c^{m,m',l,l'}_{n,n',k,k'}(q_1,q_2) \cdot W_{m,l} W_{m',l'} \Big|_\Delta \right), \tag{2-28}$$

$$[P_{n'}, P_n] = \Delta_* \begin{cases} 0 & \textit{if } \operatorname{sign}(n) = \operatorname{sign}(n'), \\ \delta^0_{n+n'} n[n]_{q_1}[n]_{q_2}[r]_{q^n} \cdot \operatorname{proj}^*_{\mathcal{M}} & \textit{if } n' < 0 < n, \end{cases} \tag{2-29}$$

$$[W_{n',k'}, P_{\pm n}] = \Delta_*(\pm [n]_{q_1}[n]_{q_2}[k']_{q^n} q^{n(r-k')\delta^+_\pm} \cdot W_{\pm n+n',k'}), \tag{2-30}$$

where the coefficients $c^{m,m',l,l'}_{n,n',k,k'}(q_1,q_2) \in K_S$ were computed algorithmically in [17]*. They are certain universal symmetric Laurent polynomials in q_1 and q_2.*

Indeed, we show in [17, Theorem 3.13] that (2-28) is equivalent to the defining relation in the deformed W–algebra $\mathcal{A}_r$ (with Δ_* replaced by $(1-q_1)(1-q_2)$). Similarly, relation (2-29) is the defining relation in the deformed Heisenberg algebra. As we explained in [17, Definition 5.2 and formulas (5.20)–(5.21)] and proved in [16, Theorem 4.15], the fact that the operators (2-20), (2-21) and (2-23) satisfy the relations in Theorem 2.10 is precisely what we mean when we say that the deformed W–algebra $\mathcal{A}_r$ and the deformed Heisenberg algebra act on the groups $K_{\mathcal{M}}$.

2.11 Let us consider the operators of Section 2.5 and form the generating series

$$L_n(y) = \sum_{k=1}^{\infty} \frac{L_{n,k}}{(-y)^k} \quad \text{and} \quad U_n(y) = \sum_{k=1}^{\infty} \frac{U_{n,k}}{(-y)^k}. \tag{2-31}$$

In other words, these power series are considered as operators

$$K_{\mathcal{M}} \xrightarrow{L_n(y)} K_{\mathcal{M}\times S}\left[\left[\frac{1}{y}\right]\right], \quad L_n(y) = (p_+ \times p_S)_*\left(\frac{1}{1-(y/\mathcal{L}_n)} \cdot p_-^*\right),$$

$$K_{\mathcal{M}} \xrightarrow{U_n(y)} K_{\mathcal{M}\times S}\left[\left[\frac{1}{y}\right]\right], \quad U_n(y) = \frac{(-1)^{rn}u^n}{q^{(r-1)n}}(p_- \times p_S)_*\left(\frac{\mathcal{Q}^{-r}}{1-(y/\mathcal{L}_n)} \cdot p_+^*\right).$$

We will also consider the operators

$$E(y)\colon K_{\mathcal{M}} \xrightarrow{\text{pullback}} K_{\mathcal{M}\times S} \xrightarrow{\text{multiplication by } \wedge^\bullet(\mathcal{U}/y)} K_{\mathcal{M}\times S}\left[\left[\frac{1}{y}\right]\right].$$

Furthermore, we will consider the generating series

$$L(x,y) = 1 + \sum_{n=1}^{\infty} L_n(y)x^n \quad \text{and} \quad U(x,y) = 1 + \sum_{n=1}^{\infty} \frac{U_n(y)}{x^n}, \tag{2-32}$$

and also set

$$W_k(x) = \sum_{n=-\infty}^{\infty} \frac{W_{n,k}}{x^n}, \tag{2-33}$$

$$W(x,y) = 1 + \sum_{k=1}^{\infty} \frac{W_k(x)}{y^k}. \tag{2-34}$$

The definition of the W–algebra generators in (2-20) is equivalent to

$$W(x, yD_x) = L(x, yD_x)E(yD_x)U(xq, yD_x)|_\Delta, \tag{2-35}$$

where D_x is the q–difference operator in the variable x, ie $D_x(f(x)) = f(xq)$. In formula (2-35), we place all powers of D_x to the right (resp. left) of all powers of x when writing down the power series $L(x, yD_x)$ (resp. $U(xq, yD_x)$). In terms of generating series, formula (2-30) reads

$$[W_k(x), P_{\pm n}] = \Delta_*\bigl(\pm[n]_{q_1}[n]_{q_2}[k]_{q^n} q^{n(r-k)\delta^+_\pm} \cdot x^{\pm n} W_k(x)\bigr). \tag{2-36}$$

2.12 Given a rational function $F(z)$, whose set of simple poles is partitioned into two disjoint sets $\mathcal{P}_1 \sqcup \mathcal{P}_2$ (which may be empty), we will write

$$\int_{\mathcal{P}_1 \prec z \prec \mathcal{P}_2} F(z) = \sum_{c \in \mathcal{P}_1} \operatorname*{Res}_{z=c} \frac{F(z)}{z} = -\sum_{c \in \mathcal{P}_2} \operatorname*{Res}_{z=c} \frac{F(z)}{z}. \tag{2-37}$$

The first equality is a definition, and the second equality is the residue theorem. If $F(z_1, \dots, z_n)$ is a rational function with simple poles of the form $z_i = c$ and $z_i = \gamma z_j$ for various $c \in \mathcal{P}_1 \sqcup \mathcal{P}_2$ and various scalars γ in some set $\mathcal{Q}$, then we set

$$\int_{\mathcal{P}_1 \prec z_1 \prec \dots \prec z_n \prec \mathcal{P}_2} F(z_1, \dots, z_n) \tag{2-38}$$

as the result of the n–step process which starts with $F(z_1, \dots, z_n)/z_1 \cdots z_n$, and at the i^{th} step replaces a rational function in $z_i, \dots, z_n$ by the sum of its residues of the form $z_i = c\gamma_1 \cdots \gamma_{i-1}$ for various $c \in \mathcal{P}_1$ and $\gamma_1, \dots, \gamma_{i-1} \in \mathcal{Q} \cup \{1\}$. Just like in (2-37), the residue theorem implies that the answer is the same as $(-1)^n$ times the result of the n–step process which starts with $F(z_1, \dots, z_n)/z_1 \cdots z_n$, and at the i^{th} step replaces a rational function in $z_1, \dots, z_{n+1-i}$ by the sum of its residues of the form $z_{n+1-i} = c\gamma_1 \cdots \gamma_{i-1}$ for various $c \in \mathcal{P}_2$ and $\gamma_1, \dots, \gamma_{i-1} \in \mathcal{Q} \cup \{1\}$.

Proposition 2.13 [17, following the proof of Proposition 5.12] *We have the following formulas for the maps* (2-13):

(2-39) $$(p_+ \times p_S)_* r(\mathcal{L}_1, \dots, \mathcal{L}_n) = \int_{\{0,\infty\} \sqcup \mathcal{P} \prec z_n \prec \dots \prec z_1 \prec \mathcal{U}} \frac{r(z_1, \dots, z_n) \prod_{i=1}^n \wedge^\bullet\left(\frac{z_i q}{\mathcal{U}}\right)}{\left(1 - \frac{z_2 q}{z_1}\right) \dots \left(1 - \frac{z_n q}{z_{n-1}}\right) \prod_{1 \le i < j \le n} \zeta\left(\frac{z_j}{z_i}\right)},$$

(2-40) $$(p_- \times p_S)_* r(\mathcal{L}_1, \dots, \mathcal{L}_n) = \int_{\mathcal{U} \prec z_n \prec \dots \prec z_1 \prec \{0,\infty\} \sqcup \mathcal{P}} \frac{r(z_1, \dots, z_n) \prod_{i=1}^n \wedge^\bullet\left(-\frac{\mathcal{U}}{z_i}\right)}{\left(1 - \frac{z_2 q}{z_1}\right) \dots \left(1 - \frac{z_n q}{z_{n-1}}\right) \prod_{1 \le i < j \le n} \zeta\left(\frac{z_j}{z_i}\right)},$$

where

$$\zeta(x) = \frac{(1 - x q_1)(1 - x q_2)}{(1 - x)(1 - x q)} \in K_S(x)$$

and $r(z_1, \dots, z_n)$ *is a rational function with coefficients in* $(p_\pm \times p_S)^*(K_{\mathcal{M} \times S})$ *whose poles are all of the form* $z_i = c$, *where* $c \in \{0, \infty\} \sqcup \mathcal{P}$ *for some finite set* $\mathcal{P}$.

Note that the integrands in (2-39)–(2-40) have poles when z_i equals $q^{1 \text{ or } 0}$ times one of the Chern roots of $\mathcal{U}$. Thus, the location of the symbol $\mathcal{U}$ in the subscripts of the integrals (2-39)–(2-40) indicates whether these poles are thought to lie in the set $\mathcal{P}_1$ or $\mathcal{P}_2$ for the sake of the notation (2-37).

3 Computing the Ext operator

3.1 To properly define the Ext operator (1-6), note that the complex $\mathcal{E}_m$ of (1-4) can be written as a difference $\mathcal{V}_1 - \mathcal{V}_2$ of vector bundles. Then we define

(3-1) $$\wedge^\bullet\left(\frac{\mathcal{E}_m}{t}\right) = \frac{\wedge^\bullet\left(\frac{\mathcal{V}_1}{t}\right)}{\wedge^\bullet\left(\frac{\mathcal{V}_2}{t}\right)} = \frac{\sum_{k=0}^{\operatorname{rank} \mathcal{V}_1} (-t)^{-k} [\wedge^k \mathcal{V}_1]}{\sum_{k=0}^{\operatorname{rank} \mathcal{V}_2} (-t)^{-k} [\wedge^k \mathcal{V}_2]}$$

and interpret it as a rational function in t, with coefficients in $K_{\mathcal{M} \times \mathcal{M}'}$. Strictly speaking, the object $\wedge^\bullet \mathcal{E}_m$ in (1-6) refers to the specialization of this rational function at $t = 1$. If this specialization is not well-defined, ie if

$$\sum_{k=0}^{\operatorname{rank}\mathcal{V}_2} (-1)^k [\wedge^k \mathcal{V}_2]$$

is not a unit in $K_{\mathcal{M}\times\mathcal{M}'}$, then we employ the following artifice: replace m by m/t in formulas (1-11), (1-12), (1-17) and throughout the current section. Once one does this, our main Theorems 1.7, 1.10 and Corollary 1.11 will be equalities of operator-valued rational functions in t. Moreover, we will often use the notation

$$\wedge^\bullet\left(\frac{t}{\mathcal{U}}\right) \quad \text{instead of} \quad \wedge^\bullet(\mathcal{U}^\vee t)$$

for any coherent sheaf $\mathcal{U}$ (all our coherent sheaves have finite projective dimension).

3.2 The main goal of the present section is to compute the commutation relations between the Ext operator $A_m\colon K_{\mathcal{M}'} \to K_{\mathcal{M}}$ of (1-6) and the operators

$$W_{n,k},\, P_{\pm n'}\colon K_{\mathcal{M}} \to K_{\mathcal{M}\times S} \tag{3-2}$$

of (2-20), (2-21) and (2-23) for all $n \in \mathbb{Z}$ and $n', k \in \mathbb{N}$. One must be careful what one means by "commutation relation". While the operator

$$P_{\pm n} A_m \quad \text{unambiguously refers to} \quad K_{\mathcal{M}'} \xrightarrow{A_m} K_{\mathcal{M}} \xrightarrow{P_{\pm n}} K_{\mathcal{M}\times S},$$

$$A_m P_{\pm n} \quad \text{henceforth refers to} \quad K_{\mathcal{M}'} \xrightarrow{P_{\pm n}} K_{\mathcal{M}'\times S} \xrightarrow{A_m\times \mathrm{Id}_S} K_{\mathcal{M}\times S},$$

and analogously for $W_{n,k}$ instead of $P_{\pm n}$. As opposed to the operators (3-2), the operator A_m acts nontrivially between all components of the moduli space

$$A_m\big|_c^{c'}\colon K_{\mathcal{M}_{c'}} \to K_{\mathcal{M}_c}. \tag{3-3}$$

In principle, the moduli spaces of sheaves in the domain and codomain can correspond to different choices of first Chern class and stability condition, but we always require them to have the same rank r. Therefore, there are two universal sheaves

$$\begin{array}{ccc} \mathcal{U} & & \mathcal{U}' \\ \downarrow & \text{and} & \downarrow \\ \mathcal{M}\times S & & \mathcal{M}'\times S \end{array}$$

of the same rank r, where $\mathcal{M}$ (resp. $\mathcal{M}'$) is the union of the moduli spaces that appear in the codomain (resp. domain) of (3-3). The determinants of these universal sheaves are denoted by u and u', respectively, as in (1-10).

3.3 We must explain how to make sense of the symbols q, m, γ in (1-12), (1-14) and (1-15). In the language of correspondences from Section 2.1, the operators

$$K_{\mathcal{M}'} \xrightarrow{z} K_{\mathcal{M}\times S}$$

studied in the present paper (such as the compositions $W_{n,k} A_m$ or $P_{\pm n} A_m$ that appear in (1-12), (1-14) and (1-15)) arise from K–theory classes Γ on $\mathcal{M} \times \mathcal{M}' \times S$. Then the product qz refers to the operator corresponding to the class $\mathrm{proj}_S^*(q) \cdot \Gamma$, while the product γz refers to the operator corresponding to the class

$$\mathrm{proj}_S^* \left(\frac{m}{q}\right)^r \cdot \frac{\mathrm{proj}_{\mathcal{M}\times S}^*(\det \mathcal{U})}{\mathrm{proj}_{\mathcal{M}'\times S}^*(\det \mathcal{U}')} \cdot \Gamma,$$

where $\mathcal{M} \times \mathcal{M}' \times S \xrightarrow{\mathrm{proj}_{\mathcal{M}\times S},\ \mathrm{proj}_{\mathcal{M}'\times S},\ \mathrm{proj}_S} \mathcal{M} \times S,\ \mathcal{M}' \times S,\ S$ are the projections.

Proposition 3.4 *We have the equality of correspondences $K_{\mathcal{M}_{c\pm n}} \to K_{\mathcal{M}_c \times S}$*

$$P_{\pm n} \cdot (\det \mathcal{U}_{c\pm n}) = (\det \mathcal{U}_c) \cdot P_{\pm n} \tag{3-4}$$

for all $c \in \mathbb{Z}$. Formula (3-4) *also holds with $P_{\pm n}$ replaced by $W_{n,k}$ or $H_{\pm n}$.*

Equation (3-4) is best restated in the language of correspondences from Section 2.1. In these terms, $P_{\pm n}$ is given by a K–theory class supported on the locus

$$\mathfrak{C} = \{(\mathcal{F}_{c+n} \subset_{nx} \mathcal{F}_c) \text{ for some } x \in S\} \subset \mathcal{M}_{c+n} \times \mathcal{M}_c \times S,$$

where $\mathcal{F}' \subset_{nx} \mathcal{F}$ means that $\mathcal{F}' \subset \mathcal{F}$ and that $\mathcal{F}/\mathcal{F}'$ is a length n sheaf supported at x. Then (3-4) merely states that the universal sheaves $\mathcal{U}_{c+n}$ and $\mathcal{U}_c$ have isomorphic determinants when restricted to $\mathfrak{C}$. This is just the version "in families" of the well-known statement that a codimension-2 modification of a torsion-free sheaf does not change its determinant. As a consequence of Proposition 3.4, γ of (1-11) will behave just like a constant in all our computations, ie it will not matter where we insert γ in any product of operators among $P_{\pm n}$, $H_{\pm n}$ and $W_{n,k}$.

3.5 Our main intersection-theoretic computation is the following:

Lemma 3.6 *We have the following relations involving the* **Ext** *operator A_m*

$$A_m(H_{-n} - H_{-n+1}) = \gamma^n (H_{-n} - H_{-n+1}) A_m, \tag{3-5}$$

$$A_m(H_n - H_{n-1}\gamma^{-1}) = (H_n \gamma^{-n} - H_{n-1} q^r \gamma^{-n+1}) A_m \tag{3-6}$$

for all $n \in \mathbb{N}$. (Recall that $H_0 = \mathrm{proj}_1^*$, where $\mathcal{M} \times S \xrightarrow{\mathrm{proj}_1} \mathcal{M}$ is the usual projection.)

Proof Consider the following diagrams of spaces and arrows, for all $c, c' \in \mathbb{Z}$:

$$\begin{tikzcd}
 & & \mathcal{M}_c \times S \times \mathcal{M}_{c'} \arrow[dddll, bend right, "\pi_1 \times \mathrm{Id}_S"'] \arrow[dddrr, bend left, "\pi_2"] & & \\
 & & \mathcal{M}_c \times \mathfrak{Z}^\bullet_{c'+n,c'} \arrow[u, "\mathrm{Id} \times p_S \times p_-"] \arrow[dl, "\mathrm{Id} \times p_+ \times p_S"'] \arrow[dr] & & \\
 & \mathcal{M}_c \times \mathcal{M}_{c'+n} \times S \arrow[dl] \arrow[dr] & & \mathfrak{Z}^\bullet_{c'+n,c'} \arrow[dl, "p_+ \times p_S"'] \arrow[dr, "p_-"] & \\
\mathcal{M}_c \times S & & \mathcal{M}_{c'+n} \times S & & \mathcal{M}_{c'}
\end{tikzcd} \tag{3-7}$$

$$\begin{tikzcd}
 & & \mathcal{M}_c \times S \times \mathcal{M}_{c'} \arrow[dddll, bend right, "\pi'_1 \times \mathrm{Id}_S"'] \arrow[dddrr, bend left, "\pi'_2"] & & \\
 & & \mathfrak{Z}^\bullet_{c,c-n} \times \mathcal{M}_{c'} \arrow[u, "p'_+ \times p'_S \times \mathrm{Id}"'] \arrow[dl] \arrow[dr, "p'_- \times \mathrm{Id}"] & & \\
 & \mathfrak{Z}^\bullet_{c,c-n} \arrow[dl, "p'_+ \times p'_S"'] \arrow[dr, "p'_-"] & & \mathcal{M}_{c-n} \times \mathcal{M}_{c'} \arrow[dl] \arrow[dr] & \\
\mathcal{M}_c \times S & & \mathcal{M}_{c-n} & & \mathcal{M}_{c'}
\end{tikzcd} \tag{3-8}$$

Recall that $H_{-n} = (p_+ \times p_S)_* p_-^*$, in the notation of (2-13). Then the rule for composition of correspondences in (2-2) gives us the formulas

$$A_m H_{-n} = (\pi_1 \times \mathrm{Id}_S)_*(\Upsilon_n \cdot \pi_2^*), \tag{3-9}$$

$$H_{-n} A_m = (\pi'_1 \times \mathrm{Id}_S)_*(\Upsilon'_n \cdot \pi_2'^*), \tag{3-10}$$

where, in the notation of (3-7) and (3-8),

$$\Upsilon_n = (\mathrm{Id} \times p_S \times p_-)_*\big[\wedge^\bullet\big((\mathrm{Id} \times p_+)^* \mathcal{E}_m\big)\big], \tag{3-11}$$

$$\Upsilon'_n = (p'_+ \times p'_S \times \mathrm{Id})_*\big[\wedge^\bullet\big((p'_- \times \mathrm{Id})^* \mathcal{E}_m\big)\big] \tag{3-12}$$

are certain classes on $\mathcal{M}_c \times S \times \mathcal{M}_{c'}$, which we will now compute.

Claim 3.7 *In K–theory we have the equalities*

$$(\mathrm{Id} \times p_+)^* \mathcal{E}_m = (\mathrm{Id} \times p_-)^* \mathcal{E}_m + \Big(\frac{1}{\mathcal{L}_1} + \cdots + \frac{1}{\mathcal{L}_n}\Big)(\mathrm{Id} \times p_S)^* \Big(\frac{\mathcal{U} m}{q}\Big) \tag{3-13}$$

on $\mathcal{M}_c \times \mathfrak{Z}^\bullet_{c'+n,c'}$, where $\mathcal{U}$ denotes the universal sheaf on $\mathcal{M}_c \times S$, and

$$(p'_- \times \mathrm{Id})^* \mathcal{E}_m = (p'_+ \times \mathrm{Id})^* \mathcal{E}_m - (\mathcal{L}_1 + \cdots + \mathcal{L}_n)(p'_S \times \mathrm{Id})^* (\mathcal{U}'^\vee m) \tag{3-14}$$

on $\mathfrak{Z}^\bullet_{c,c-n} \times \mathcal{M}_{c'}$, where $\mathcal{U}'$ denotes the universal sheaf on $\mathcal{M}_{c'} \times S$.

Proof To prove (3-13), consider the diagram

$$
\begin{tikzcd}
 & \mathcal{M}_c \times \mathfrak{Z}^\bullet_{c'+n,c'} \times S \arrow[dl, "\mathrm{Id} \times p_+ \times \mathrm{Id}_S"'] \arrow[d, "\rho"] \arrow[dr, "\mathrm{Id} \times p_- \times \mathrm{Id}_S"] & \\
\mathcal{M}_c \times \mathcal{M}_{c'+n} \times S \arrow[d, "\rho"] & \mathcal{M}_c \times \mathfrak{Z}^\bullet_{c'+n,c'} \arrow[dl, "\mathrm{Id} \times p_+"'] \arrow[dr, "\mathrm{Id} \times p_-"] & \mathcal{M}_c \times \mathcal{M}_{c'} \times S \arrow[d, "\rho"] \\
\mathcal{M}_c \times \mathcal{M}_{c'+n} & & \mathcal{M}_c \times \mathcal{M}_{c'}
\end{tikzcd}
\tag{3-15}
$$

where the vertical maps are the natural projections (we use the notation ρ for all of them). We have the short exact sequence of sheaves over $\mathfrak{Z}^\bullet_{c'+n,c'} \times S$

$$0 \to \mathcal{U}'_+ \to \mathcal{U}'_- \to \Gamma_*(\mathcal{L}_1 \text{“}\oplus\text{”} \cdots \text{“}\oplus\text{”} \mathcal{L}_n) \to 0, \tag{3-16}$$

where $\mathcal{U}'_\pm = (p^*_\pm \times \mathrm{Id}_S)(\text{universal sheaf})$, while $\mathcal{L}_1, \dots, \mathcal{L}_n$ denote the tautological line bundles on $\mathfrak{Z}^\bullet_{c'+n,c'}$ that were defined in (2-14), and

$$\Gamma \colon \mathfrak{Z}^\bullet_{c'+n,c'} \to \mathfrak{Z}^\bullet_{c'+n,c'} \times S \tag{3-17}$$

is the graph of the map p_S. The notation “$\oplus$” in (3-16) refers to a coherent sheaf which is filtered by the line bundles $\mathcal{L}_1, \dots, \mathcal{L}_n$; since we work in K–theory, we henceforth make no distinction between this coherent sheaf and its associated graded object. We may also pull back the short exact sequence (3-16) to $\mathcal{M}_c \times \mathfrak{Z}^\bullet_{c'+n,c'} \times S$. Now apply the functor $\mathrm{R}\mathscr{H}\!om(-, \mathcal{U} \otimes m)$ to the short exact sequence (3-16), where $\mathcal{U}$ is the universal sheaf pulled back from $\mathcal{M}_c \times S$:

$$\mathrm{R}\mathscr{H}\!om(\mathcal{U}'_+, \mathcal{U} \otimes m) = \mathrm{R}\mathscr{H}\!om(\mathcal{U}'_-, \mathcal{U} \otimes m) - \sum_{i=1}^n \frac{1}{\mathcal{L}_i} \mathrm{R}\mathscr{H}\!om(\mathcal{O}_\Gamma, \mathcal{U} \otimes m).$$

Now recall that the line bundles $\mathcal{L}_i$ come from $\mathfrak{Z}^\bullet_{c'+n,c'}$, and so they are unaffected by the derived pushforward map ρ_*,

$$\rho_* \mathrm{R}\mathscr{H}\!om(\mathcal{U}'_+, \mathcal{U} \otimes m) = \rho_* \mathrm{R}\mathscr{H}\!om(\mathcal{U}'_-, \mathcal{U} \otimes m) - \sum_{i=1}^n \frac{1}{\mathcal{L}_i} \rho_* \mathrm{R}\mathscr{H}\!om(\mathcal{O}_\Gamma, \mathcal{U} \otimes m).$$

Recalling (1-5), the formula above reads

$$(\mathrm{Id} \times p_+)^* \mathcal{E}_m = (\mathrm{Id} \times p_-)^* \mathcal{E}_m + \sum_{i=1}^n \frac{1}{\mathcal{L}_i} \rho_* \mathrm{R}\mathscr{H}\!om(\mathcal{O}_\Gamma, \mathcal{U} \otimes m). \tag{3-18}$$

Then (3-13) follows from the fact that

$$\rho_* \mathrm{R}\mathscr{H}\!om(\mathcal{O}_\Gamma, \mathcal{U} \otimes m) = \overbrace{\rho_* \circ \Gamma_*}^{\mathrm{Id}} \big(\mathrm{R}\mathscr{H}\!om(\mathcal{O}, \Gamma^!(\mathcal{U} \otimes m))\big) = \mathcal{U}m|_\Gamma \otimes \Gamma^! \mathcal{O}. \tag{3-19}$$

(The first equality is coherent duality, and the second equality holds for any closed embedding Γ.) The right-hand side of (3-19) matches $(\mathrm{Id} \times p_S)^*(\mathcal{U}m/q)$ because the map $\Gamma\colon \mathfrak{Z}_n^\bullet \to \mathfrak{Z}_n^\bullet \times S$ is obtained by base change from the diagonal map $S \to S \times S$, and the ratio of dualizing objects on S and $S \times S$ is precisely $q = [\omega_S]$.

As for (3-14), consider the diagram

$$\begin{tikzcd}
 & \mathfrak{Z}^\bullet_{c,c-n} \times \mathcal{M}_{c'} \times S \arrow[dl, "p'_+ \times \mathrm{Id} \times \mathrm{Id}_S"'] \arrow[dd, "\rho"] \arrow[dr, "p'_- \times \mathrm{Id} \times \mathrm{Id}_S"] & \\
\mathcal{M}_c \times \mathcal{M}_{c'} \times S \arrow[dd, "\rho"] & & \mathcal{M}_{c-n} \times \mathcal{M}_{c'} \times S \arrow[dd, "\rho"] \\
 & \mathfrak{Z}^\bullet_{c,c-n} \times \mathcal{M}_{c'} \arrow[dl, "p'_+ \times \mathrm{Id}"'] \arrow[dr, "p'_- \times \mathrm{Id}"] & \\
\mathcal{M}_c \times \mathcal{M}_{c'} & & \mathcal{M}_{c-n} \times \mathcal{M}_{c'}
\end{tikzcd} \tag{3-20}$$

and consider the following analogue of (3-16):

$$0 \to \mathcal{U}_+ \to \mathcal{U}_- \to \Gamma'_*(\mathcal{L}_1 \text{“}\oplus\text{”} \cdots \text{“}\oplus\text{”}\, \mathcal{L}_n) \to 0,$$

where $\mathcal{U}_\pm = ({p'}^*_\pm \times \mathrm{Id}_S)(\mathcal{U})$, and Γ' denotes the graph of the map $p_S\colon \mathfrak{Z}^\bullet_{c,c-n} \to S$. Let us apply the functor $\mathrm{R}\mathscr{H}\!om(\mathcal{U}', - \otimes m)$ to the short exact sequence above:

$$\mathrm{R}\mathscr{H}\!om(\mathcal{U}', \mathcal{U}_- \otimes m) = \mathrm{R}\mathscr{H}\!om(\mathcal{U}', \mathcal{U}_+ \otimes m) + \sum_{i=1}^n \mathcal{L}_i \otimes \mathrm{R}\mathscr{H}\!om(\mathcal{U}', \mathcal{O}_{\Gamma'} \otimes m).$$

Let us apply ρ_* to the equality above, and recall the definition of $\mathcal{E}_m$ in (1-5):

$$(p'_- \times \mathrm{Id})^* \mathcal{E}_m = (p'_+ \times \mathrm{Id})^* \mathcal{E}_m - \sum_{i=1}^n \mathcal{L}_i \otimes \rho_* \mathrm{R}\mathscr{H}\!om(\mathcal{U}', \mathcal{O}_{\Gamma'} \otimes m).$$

By adjunction, we have

$$\rho_* \mathrm{R}\mathscr{H}\!om(\mathcal{U}', \mathcal{O}_{\Gamma'} \otimes m) = \overbrace{\rho_* \circ \Gamma'_*}^{\mathrm{Id}} \mathrm{R}\mathscr{H}\!om(\mathcal{U}'|_{\Gamma'}, {p'_S}^* m) = ({\mathcal{U}'}^\vee m)|_{\Gamma'}. \qquad \square$$

Armed with (3-13) and (3-14), we may rewrite (3-11) and (3-12) as

$$\Upsilon_n = [\wedge^\bullet \mathcal{E}_m] \cdot (\mathrm{Id} \times p_S \times p_-)_* \left[\bigotimes_{i=1}^n \wedge^\bullet \left(\frac{\mathcal{U}m}{\mathcal{L}_i q} \right) \right],$$

$$\Upsilon'_n = [\wedge^\bullet \mathcal{E}_m] \cdot (p'_+ \times p'_S \times \mathrm{Id})_* \left[\bigotimes_{i=1}^n \wedge^\bullet \left(-\frac{\mathcal{L}_i m}{\mathcal{U}'} \right) \right].$$

Henceforth, "$\mathcal{U}, \mathcal{U}'$" in the subscript of the integrals are simply shorthand for "the set of Chern roots of $\mathcal{U}, \mathcal{U}'$", respectively, and Proposition 2.13 implies

$$\Upsilon_n = [\wedge^\bullet \mathcal{E}_m] \int_{\mathcal{U}' \prec z_n \prec \cdots \prec z_1 \prec \{0,\infty\} \sqcup \mathcal{U}} \frac{\prod_{i=1}^n \frac{\wedge^\bullet(\mathcal{U}m/(z_i q))}{\wedge^\bullet(\mathcal{U}'/z_i)}}{\prod_{i=1}^{n-1} (1 - (q z_{i+1}/z_i)) \prod_{i<j} \zeta(z_j/z_i)}, \tag{3-21}$$

$$\Upsilon'_n = [\wedge^\bullet \mathcal{E}_m] \int_{\{0,\infty\} \sqcup \mathcal{U}' \prec z_n \prec \cdots \prec z_1 \prec \mathcal{U}} \frac{\prod_{i=1}^n \frac{\wedge^\bullet(z_i q/\mathcal{U})}{\wedge^\bullet(z_i m/\mathcal{U}')}}{\prod_{i=1}^{n-1} (1 - (q z_{i+1}/z_i)) \prod_{i<j} \zeta(z_j/z_i)}. \tag{3-22}$$

Consider the rational function with coefficients in $K_{\mathcal{M}_c \times S \times \mathcal{M}_{c'}}$ given by

$$I_n(z_1, \dots, z_n) = \frac{\prod_{i=1}^n \frac{\wedge^\bullet(\mathcal{U}m/(z_i q))}{\wedge^\bullet(\mathcal{U}'/z_i)}}{\prod_{i=1}^{n-1} (1 - (q z_{i+1}/z_i)) \prod_{i<j} \zeta(z_j/z_i)}. \tag{3-23}$$

One may then rewrite (3-21) and (3-22) as

$$\Upsilon_n = [\wedge^\bullet \mathcal{E}_m] \int_{\mathcal{U}' \prec z_n \prec \cdots \prec z_1 \prec \{0,\infty\} \sqcup \mathcal{U}} I_n(z_1, \dots, z_n),$$

$$\Upsilon'_n = [\wedge^\bullet \mathcal{E}_m] \int_{\{0,\infty\} \sqcup \mathcal{U}' \prec z_n \prec \cdots \prec z_1 \prec \mathcal{U}} I_n(z_1 m, \dots, z_n m) \cdot \gamma^{-n}.$$

Changing the variables $z_i \mapsto z_i/m$ in the second formula, we conclude that

$$\Upsilon_n - \Upsilon'_n \cdot \gamma^n = [\wedge^\bullet \mathcal{E}_m] \left[\int_{\mathcal{U}' \prec z_n \prec \cdots \prec z_1 \prec \{0,\infty\} \sqcup \mathcal{U}} I_n - \int_{\{0,\infty\} \sqcup \mathcal{U}' \prec z_n \prec \cdots \prec z_1 \prec \mathcal{U}} I_n \right]. \tag{3-24}$$

The only difference between the two integrals is the location of the poles $\{0, \infty\}$ with respect to the variables $z_1, \dots, z_n$. Therefore, we conclude that the difference above picks up the residues at 0 and ∞ in the various variables. However, all such residues vanish, except for

$$\operatorname*{Res}_{z_1=\infty} \frac{I_n(z_1,\dots,z_n)}{z_1} = -I_{n-1}(z_2,\dots,z_n), \tag{3-25}$$

$$\operatorname*{Res}_{z_n=0} \frac{I_n(z_1,\dots,z_n)}{z_n} = \gamma \cdot I_{n-1}(z_1,\dots,z_{n-1}). \tag{3-26}$$

Therefore, formula (3-24) implies that

$$\Upsilon_n - \Upsilon'_n \cdot \gamma^n = \Upsilon_{n-1} - \Upsilon'_{n-1} \cdot \gamma^n \tag{3-27}$$

which, as an equality of classes on $\mathcal{M}_c \times S \times \mathcal{M}_{c'}$, precisely encodes (3-5). Let us run the analogous computation for (3-6) (we will recycle all of our notation):

$$\begin{tikzcd}
& & \mathcal{M}_c \times S \times \mathcal{M}_{c'} \arrow[dddll, bend right, "\pi_1\times \mathrm{Id}_S"'] \arrow[dddrr, bend left, "\pi_2"] & & \\
& & \mathcal{M}_c \times \mathfrak{Z}^\bullet_{c',c'-n} \arrow[u, "\mathrm{Id}\times p_S\times p_+"] \arrow[dl, "\mathrm{Id}\times p_-\times p_S"'] \arrow[dr] & & \\
& \mathcal{M}_c\times\mathcal{M}_{c'-n}\times S \arrow[dl] \arrow[dr] & & \mathfrak{Z}^\bullet_{c',c'-n} \arrow[dl, "p_-\times p_S"'] \arrow[dr, "p_+"] & \\
\mathcal{M}_c\times S & & \mathcal{M}_{c'-n}\times S & & \mathcal{M}_{c'}
\end{tikzcd} \tag{3-28}$$

$$\begin{tikzcd}
& & \mathcal{M}_c \times S \times \mathcal{M}_{c'} \arrow[dddll, bend right, "\pi'_1\times \mathrm{Id}_S"'] \arrow[dddrr, bend left, "\pi'_2"] & & \\
& & \mathfrak{Z}^\bullet_{c+n,c} \times \mathcal{M}_{c'} \arrow[u, "p'_-\times p'_S\times \mathrm{Id}"] \arrow[dl] \arrow[dr, "p'_+\times \mathrm{Id}"] & & \\
& \mathfrak{Z}^\bullet_{c+n,c} \arrow[dl, "p'_-\times p'_S"'] \arrow[dr, "p'_+"] & & \mathcal{M}_{c+n}\times\mathcal{M}_{c'} \arrow[dl] \arrow[dr] & \\
\mathcal{M}_c\times S & & \mathcal{M}_{c+n} & & \mathcal{M}_{c'}
\end{tikzcd} \tag{3-29}$$

Recall that $H_n = (-1)^{rn} u^n (p_- \times p_S)_* (\mathcal{Q}^{-r} \cdot p_+^*)$, in the notation of (2-13). Then the rule for composition of correspondences in (2-2) gives us

$$A_m H_n = (\pi_1 \times \mathrm{Id}_S)_* (\Upsilon_n \cdot \pi_2^*), \tag{3-30}$$

$$H_n A_m = (\pi'_1 \times \mathrm{Id}_S)_* (\Upsilon'_n \cdot \pi_2'^*), \tag{3-31}$$

where

$$\Upsilon_n = (-1)^{rn} u'^n (\mathrm{Id} \times p_S \times p_+)_* \big[\mathcal{Q}^{-r} \cdot \wedge^\bullet \big((\mathrm{Id} \times p_-)^* \mathcal{E}_m\big)\big], \tag{3-32}$$

$$\Upsilon'_n = (-1)^{rn} u^n (p'_- \times p'_S \times \mathrm{Id})_* \big[\mathcal{Q}^{-r} \cdot \wedge^\bullet \big((p'_+ \times \mathrm{Id})^* \mathcal{E}_m\big)\big] \tag{3-33}$$

are certain classes on $\mathcal{M}_c \times S \times \mathcal{M}_{c'}$. As a consequence of (3-13) and (3-14), which continue to hold as stated in the new setup, we may rewrite (3-32) and (3-33) as

$$\Upsilon_n = (-1)^{rn} u'^n [\wedge^\bullet \mathcal{E}_m] (\mathrm{Id} \times p_S \times p_+)_* \left[\mathcal{Q}^{-r} \bigotimes_{i=1}^n \wedge^\bullet \left(-\frac{\mathcal{U} m}{\mathcal{L}_i q}\right)\right],$$

$$\Upsilon'_n = (-1)^{rn} u^n [\wedge^\bullet \mathcal{E}_m] (p'_- \times p'_S \times \mathrm{Id})_* \left[\mathcal{Q}^{-r} \bigotimes_{i=1}^n \wedge^\bullet \left(\frac{\mathcal{L}_i m}{\mathcal{U}'}\right)\right].$$

Therefore, Proposition 2.13 implies

$$\Upsilon_n = [\wedge^\bullet \mathcal{E}_m] \int_{\{0,\infty\} \sqcup \mathcal{U} \prec z_n \prec \cdots \prec z_1 \prec \mathcal{U}'} \frac{(-1)^{rn} u'^n z_1^{-r} \ldots z_n^{-r} \prod_{i=1}^n \dfrac{\wedge^\bullet(z_i q / \mathcal{U}')}{\wedge^\bullet(\mathcal{U} m / (z_i q))}}{\prod_{i=1}^{n-1} (1 - (q z_{i+1}/z_i)) \prod_{i<j} \zeta(z_j/z_i)}, \tag{3-34}$$

$$\Upsilon'_n = [\wedge^\bullet \mathcal{E}_m] \int_{\mathcal{U} \prec z_n \prec \cdots \prec z_1 \prec \{0,\infty\} \sqcup \mathcal{U}'} \frac{(-1)^{rn} u^n z_1^{-r} \ldots z_n^{-r} \prod_{i=1}^n \dfrac{\wedge^\bullet(z_i m / \mathcal{U}')}{\wedge^\bullet(\mathcal{U} / z_i)}}{\prod_{i=1}^{n-1} (1 - (q z_{i+1}/z_i)) \prod_{i<j} \zeta(z_j/z_i)}. \tag{3-35}$$

Consider the rational function with coefficients in $K_{\mathcal{M}_c \times S \times \mathcal{M}_{c'}}$ given by

$$I_n(z_1, \ldots, z_n) = \frac{q^{rn} \prod_{i=1}^n \dfrac{\wedge^\bullet(\mathcal{U}'/(z_i q))}{\wedge^\bullet(\mathcal{U} m/(z_i q))}}{\prod_{i=1}^{n-1} (1 - (q z_{i+1}/z_i)) \prod_{i<j} \zeta(z_j/z_i)} \tag{3-36}$$

One may then rewrite (3-34) and (3-35) as

$$\Upsilon_n = [\wedge^\bullet \mathcal{E}_m] \int_{\{0,\infty\} \sqcup \mathcal{U} \prec z_n \prec \cdots \prec z_1 \prec \mathcal{U}'} I_n(z_1, \ldots, z_n),$$

$$\Upsilon'_n = [\wedge^\bullet \mathcal{E}_m] \int_{\mathcal{U} \prec z_n \prec \cdots \prec z_1 \prec \{0,\infty\} \sqcup \mathcal{U}'} I_n\left(\frac{z_1 m}{q}, \ldots, \frac{z_n m}{q}\right) \cdot \gamma^n.$$

Changing the variables $z_i \mapsto z_i q/m$ in the second formula, we conclude that

$$\Upsilon_n - \Upsilon'_n \cdot \gamma^{-n} = [\wedge^\bullet \mathcal{E}_m] \left[\int_{\{0,\infty\} \sqcup \mathcal{U} \prec z_n \prec \cdots \prec z_1 \prec \mathcal{U}'} I_n - \int_{\mathcal{U} \prec z_n \prec \cdots \prec z_1 \prec \{0,\infty\} \sqcup \mathcal{U}'} I_n\right]. \tag{3-37}$$

The only difference between the two integrals is the location of the poles $\{0,\infty\}$ with respect to the variables $z_1,\dots,z_n$. Therefore, we conclude that the difference above picks up the residues at 0 and ∞ in the various variables. However, all such residues vanish, except for

$$\operatorname*{Res}_{z_n=0} \frac{I_n(z_1,\dots,z_n)}{z_n} = \gamma^{-1}\cdot I_{n-1}(z_1,\dots,z_{n-1}),$$
$$\operatorname*{Res}_{z_1=\infty} \frac{I_n(z_1,\dots,z_n)}{z_1} = -q^r\cdot I_{n-1}(z_2,\dots,z_n).$$

Therefore, formula (3-37) implies that

$$\Upsilon_n - \Upsilon'_n\cdot\gamma^{-n} = \Upsilon_{n-1}\cdot\gamma^{-1} - \Upsilon'_{n-1}\cdot q^r\gamma^{-n+1}, \tag{3-38}$$

which, as an equality of classes on $\mathcal{M}_c\times S\times\mathcal{M}_{c'}$, precisely encodes (3-6). □

3.8 We will now show how Lemma 3.6 allows us to prove Theorem 1.10.

Proof of Theorem 1.10 We will only prove (1-14), since (1-15) is analogous. We will use formulas (2-25), which say that the H operators are to the P operators as complete symmetric functions are to power sum functions. Then let us place (3-5) into a generating series that goes over all $n\in\mathbb{N}$,

$$\sum_{n=0}^{\infty} A_m H_{-n}(z^n - z^{n+1}) = \sum_{n=0}^{\infty}\big((\gamma z)^n-(\gamma z)^{n+1}\big)H_{-n}A_m. \tag{3-39}$$

If we write $H_-(z)$ for the power series (2-25) (with sign $\pm = -$), then (3-39) reads

$$A_m H_-(z)(1-z) = H_-(z\gamma)(1-\gamma z)A_m. \tag{3-40}$$

If P is an operator $K_{\mathcal{M}}\to K_{\mathcal{M}\times S}$ which commutes with two line bundles ℓ and ℓ' (in the sense of Proposition 3.4, and the discussion after it), then

$$A\exp(P)\exp(\ell')|_\Delta = \exp(P)\exp(\ell)|_\Delta A \iff AP + A\ell' = PA + \ell A. \tag{3-41}$$

(This claim uses the associativity of the operation $x,y \rightsquigarrow xy|_\Delta$, as discussed in Section 2.3.) With this in mind, formula (3-40) implies

$$A_m P_-(z) - \sum_{n=1}^{\infty}\frac{A_m}{nz^{-n}} = P_-(z\gamma)A_m - \sum_{n=1}^{\infty}\gamma^n\frac{A_m}{nz^{-n}},$$

where $P_-(z)=\sum_{n=1}^{\infty}P_{-n}/(nz^{-n})$. Extracting the coefficient of z^n yields precisely equation (1-14). □

3.9 Having proved Lemma 3.6, we will now perform the analogous computations for the commutator of A_m with the operators of Section 2.5.

Lemma 3.10 *We have the following relations involving the Ext operator A_m:*

$$A_m L_n(y) - A_m L_{n-1}(y) = L_n\left(\frac{y}{m}\right) A_m \cdot \gamma^n - L_{n-1}\left(\frac{yq}{m}\right) E\left(\frac{yq}{m}\right) A_m E(y)^{-1}\Big|_\Delta \cdot \gamma^{n-1}, \tag{3-42}$$

$$U_n\left(\frac{yq}{m}\right) A_m \cdot \gamma^{-n} - U_{n-1}\left(\frac{yq}{m}\right) A_m \cdot q\gamma^{-n+1} = A_m U_n(y) - E\left(\frac{yq}{m}\right)^{-1} A_m E(yq) U_{n-1}(yq)\Big|_\Delta \cdot q. \tag{3-43}$$

The two sides of (3-42) *and* (3-43) *map $K_{\mathcal{M}'}$ to $K_{\mathcal{M}\times S}[\![y^{-1}]\!]$. The symbol $|_\Delta$ applied to any term that involves three of the series L, E, U means that we restrict a certain operator $K_{\mathcal{M}'} \to K_{\mathcal{M}\times S\times S\times S}[\![y^{-1}]\!]$ to the small diagonal.*

Proof In order to prove (3-42), we will closely follow the proof of Lemma 3.6. With the notation therein, one needs to replace (3-11) and (3-12) by

$$\Upsilon_{n,y} = (\mathrm{Id} \times p_S \times p_-)_*\left[\frac{1}{1-(y/\mathcal{L}_n)} \wedge^\bullet \big((\mathrm{Id}\times p_+)^* \mathcal{E}_m\big)\right],$$

$$\Upsilon'_{n,y} = (p'_+ \times p'_S \times \mathrm{Id})_*\left[\frac{1}{1-(y/\mathcal{L}_n)} \wedge^\bullet \big((p'_- \times \mathrm{Id})^* \mathcal{E}_m\big)\right].$$

This has the effect of inserting

$$\left(1 - \frac{y}{z_n}\right)^{-1}$$

into the right-hand sides of formulas (3-21) and (3-22). Therefore, the function $I_n(z_1, \dots, z_n)$ defined in (3-23) should be replaced by

$$I_{n,y}(z_1, \dots, z_n) = \frac{I_n(z_1, \dots, z_n)}{1-(y/z_n)}.$$

It is easy to see that the nonzero residues of $I_{n,y}$ are

$$\operatorname*{Res}_{z_1=\infty} \frac{I_{n,y}(z_1, \dots, z_n)}{z_1} = -I_{n-1,y}(z_2, \dots, z_n),$$

$$\operatorname*{Res}_{z_n=y} \frac{I_{n,y}(z_1, \dots, z_n)}{z_n} = \frac{\wedge^\bullet(\mathcal{U}m/(yq))}{\wedge^\bullet(\mathcal{U}'/y)} \cdot \frac{I_{n-1,yq}(z_1, \dots, z_{n-1})}{\prod_{i=1}^{n-1} \zeta(y/z_i)}.$$

Therefore, the analogue of identity (3-27) is

$$\Upsilon_{n,y} - \Upsilon'_{n,y/m} \cdot \gamma^n = \Upsilon_{n-1,y} - \Upsilon'_{n-1,yq/m} \cdot \gamma^{n-1} \frac{\wedge^\bullet(\mathcal{U}m/(yq))}{\wedge^\bullet(\mathcal{U}'/y)}.$$

This equality of classes on $\mathcal{M}_c \times S \times \mathcal{M}_{c'}$ precisely underlies equality (3-42).

As for (3-43), we proceed analogously. One needs to replace (3-32) and (3-33) by

$$\Upsilon_n = \frac{(-1)^{rn} u'^n}{q^{(r-1)n}} (\mathrm{Id} \times p_S \times p_+)_* \left[\frac{\mathcal{Q}^{-r}}{1 - y/\mathcal{L}_n} \cdot \wedge^\bullet\big((\mathrm{Id} \times p_-)^* \mathcal{E}_m\big) \right],$$

$$\Upsilon'_n = \frac{(-1)^{rn} u^n}{q^{(r-1)n}} (p'_- \times p'_S \times \mathrm{Id})_* \left[\frac{\mathcal{Q}^{-r}}{1 - y/\mathcal{L}_n} \cdot \wedge^\bullet\big((p'_+ \times \mathrm{Id})^* \mathcal{E}_m\big) \right].$$

This has the effect of inserting

$$q^{n(1-r)} \left(1 - \frac{y}{z_n}\right)^{-1}$$

into the right-hand sides of formulas (3-34) and (3-35). Therefore, the function I_n defined in (3-36) should be replaced by

$$I_{n,y}(z_1, \dots, z_n) = \frac{I_n(z_1, \dots, z_n)}{q^{(r-1)n}(1 - y/z_n)}.$$

It is easy to see that the nonzero residues of $I_{n,y}$ are

$$\operatorname*{Res}_{z_n = y} \frac{I_{n,y}(z_1, \dots, z_n)}{z_n} = q \frac{\wedge^\bullet(\mathcal{U}'/(yq))}{\wedge^\bullet(\mathcal{U}m/(yq))} \cdot \frac{I_{n-1,yq}(z_1, \dots, z_{n-1})}{\prod_{i=1}^{n-1} \zeta(y/z_i)},$$

$$\operatorname*{Res}_{z_1 = \infty} \frac{I_{n,y}(z_1, \dots, z_n)}{z_1} = -q \cdot I_{n-1,y}(z_2, \dots, z_n).$$

Therefore, the analogue of identity (3-38) is

$$\Upsilon_{n,y} - \Upsilon_{n,yq/m} \cdot \gamma^{-n} = \Upsilon_{n-1,yq} \cdot q \frac{\wedge^\bullet(\mathcal{U}'/(yq))}{\wedge^\bullet(\mathcal{U}m/(yq))} - \Upsilon_{n-1,yq/m} \cdot q\gamma^{-n+1}.$$

This equality of classes on $\mathcal{M}_c \times S \times \mathcal{M}_{c'}$ precisely underlies equality (3-43). □

3.11 In all formulas below, whenever one encounters a product of several L, E, U operators, one needs to place the symbol $|_\Delta$ next to it, eg $L(\dots)E(\dots)U(\dots)|_\Delta$ as in (2-20). From now on, we will suppress the notation $|_\Delta$ from our formulas for brevity.

Proof of Theorem 1.7 In terms of the generating series (2-32), formulas (3-42) and (3-43) take the form

$$(1-x)A_m L(x,y) = L\left(x\gamma, \frac{y}{m}\right)A_m - xL\left(x\gamma, \frac{yq}{m}\right)E\left(\frac{yq}{m}\right)A_m E(y)^{-1},$$

$$U\left(x\gamma, \frac{yq}{m}\right)A_m\left(1-\frac{q}{x}\right) = A_m U(x,y) - \frac{q}{x}E\left(\frac{yq}{m}\right)^{-1} A_m E(yq)U(x,yq).$$

Change the variables $x \mapsto xq$, $y \mapsto y/q$ in the second equation, and multiply the first equation by $E(y)$ and the second equation by $E(y/m)$. Thus we obtain

$$(1-x)A_m L(x,y)E(y) = L\left(x\gamma, \frac{y}{m}\right)A_m E(y) - xL\left(x\gamma, \frac{yq}{m}\right)E\left(\frac{yq}{m}\right)A_m,$$

$$E\left(\frac{y}{m}\right)U\left(xq\gamma, \frac{y}{m}\right)A_m\left(1-\frac{1}{x}\right) = E\left(\frac{y}{m}\right)A_m U\left(xq, \frac{y}{q}\right) - \frac{1}{x}A_m E(y)U(xq,y).$$

Now let us replace the variable y by the symbol yD_x, where D_x denotes the q–difference operator $D_x(f(x)) = f(xq)$. However, we make the following prescription: in the first equation above, the D_x's are placed to the right of all x's, while in the second equation, the D_x's are placed to the left of all the x's. We thus obtain

$$(1-x)A_m L(x,yD_x)E(yD_x) = L\left(x\gamma, \frac{yD_x}{m}\right)A_m E(yD_x) - xL\left(x\gamma, \frac{yD_xq}{m}\right)E\left(\frac{yD_xq}{m}\right)A_m,$$

$$E\left(\frac{yD_x}{m}\right)U\left(xq\gamma, \frac{yD_x}{m}\right)A_m(1-x) = A_m E(yD_x)U(xq,yD_x) - E\left(\frac{yD_x}{m}\right)A_m U\left(xq, \frac{yD_x}{q}\right)x.$$

Now let us multiply the first equation on the right by $U(qx, yD_x)$ (with the D_x's placed to the left of all the x's) and the second equation on the left by $L(x\gamma, yD_x/m)$ (with the D_x's placed to the right of all the x's):

$$\begin{aligned}(1-x)A_m L(x,yD_x)E(yD_x)U(xq,yD_x) &= L\left(x\gamma, \frac{yD_x}{m}\right)A_m E(yD_x)U(xq,yD_x) \\ &\quad - xL\left(x\gamma, \frac{yD_xq}{m}\right)E\left(\frac{yD_xq}{m}\right)A_m U(xq,yD_x)\end{aligned}$$

and

$$L\left(x\gamma, \frac{yD_x}{m}\right)E\left(\frac{yD_x}{m}\right)U\left(xq\gamma, \frac{yD_x}{m}\right)A_m(1-x)$$
$$= L\left(x\gamma, \frac{yD_x}{m}\right)A_m E(yD_x)U(xq, yD_x)$$
$$- L\left(x\gamma, \frac{yD_x}{m}\right)E\left(\frac{yD_x}{m}\right)A_m U\left(xq, \frac{yD_x}{q}\right)x.$$

The two terms in the right-hand sides of the above equations are pairwise equal to each other (this is not manifestly obvious for the second term, because y differs from yq, but this is a consequence of commuting D_x past x). We conclude that

$$(1-x)A_m L(x, yD_x)E(yD_x)U(xq, yD_x)$$
$$= L\left(x\gamma, \frac{yD_x}{m}\right)E\left(\frac{yD_x}{m}\right)U\left(xq\gamma, \frac{yD_x}{m}\right)A_m(1-x).$$

Recalling the definition (2-35), this implies

$$(1-x)A_m W(x, yD_x) = W\left(x\gamma, \frac{yD_x}{m}\right)A_m(1-x).$$

Taking the coefficient of $(yD_x)^{-k}$ implies (1-12). In doing so, the right-most factor $1-x$ changes into $1-x/q^k$ due to the fact that the operators $1/D_x^k$ must pass over it. $\square$

3.12 Finally, we recall the operator $\Phi_m \colon K_{\mathcal{M}'} \to K_{\mathcal{M}}$ defined in (1-16),

$$\Phi_m = A_m \exp\left[\sum_{n=1}^{\infty} \frac{P_n}{n}\left\{\frac{(q^n-1)q^{-rn}}{[n]_{q_1}[n]_{q_2}}\right\}\right],$$

and let us translate (1-12), (1-14) and (1-15) into commutation relations involving Φ_m.

Proof of Corollary 1.11 Since P_n commutes with $P_{n'}$ for all $n, n' > 0$, (1-15) implies (1-18) when the sign is $+$. Let us now prove (1-18) when the sign is $-$. We write

$$\Phi_m = A_m \cdot \exp,$$

where exp is shorthand for

$$\exp\left[\sum_{n=1}^{\infty} \frac{P_n}{n}\left\{\frac{(q^n-1)q^{-rn}}{[n]_{q_1}[n]_{q_2}}\right\}\right].$$

Then (1-14) reads

$$\Phi_m \cdot \exp^{-1} \cdot P_{-n} - P_{-n} \cdot \Phi_m \cdot \exp^{-1} \gamma^n = \Phi_m \cdot \exp^{-1}(1-\gamma^n).$$

The relation above will establish (1-18) for $\pm = -$ once we prove that

$$[\exp^{-1}, P_{-n}] = (1 - q^{-rn}) \exp^{-1}. \tag{3-44}$$

If we take the logarithm of (3-44), it boils down to

$$\left[P_{-n}, \frac{P_n}{n}\left\{\frac{(q^n - 1)q^{-nr}}{[n]_{q_1}[n]_{q_2}}\right\}\right] = 1 - q^{-rn}. \tag{3-45}$$

Relation (3-45) is an equality of operators $K_{\mathcal{M}} \to K_{\mathcal{M}\times S}$ (the right-hand side denotes pullback multiplied by $\mathrm{proj}_S^*(1 - q^{-rn})$), and it is proved as follows. Take equality (2-29) of operators $K_{\mathcal{M}} \to K_{\mathcal{M}\times S\times S}$, multiply it by

$$\mathrm{proj}_3^*\left(\frac{1}{n}\cdot\frac{(q^n - 1)q^{-nr}}{[n]_{q_1}[n]_{q_2}}\right) \in K_{\mathcal{M}\times S\times S} \tag{3-46}$$

and then apply proj_{12*} to the result (above, we write $\mathcal{M}\times S\times S \xrightarrow{\mathrm{proj}_{12},\ \mathrm{proj}_3} \mathcal{M}\times S,\ S$ for the obvious projection maps). The outcome of this procedure is precisely (3-45).

Now let us prove (1-12) $\Longrightarrow$ (1-17). To do so, we must take formula (2-36) for $\pm = +$ (which is a priori an equality of operators $K_{\mathcal{M}} \to K_{\mathcal{M}\times S\times S}$), multiply it by (3-46) and then apply proj_{12*} to the result. The resulting equality reads

$$\left[W_k(x), \frac{P_n}{n}\left\{\frac{(q^n - 1)q^{-nr}}{[n]_{q_1}[n]_{q_2}}\right\}\right] = \frac{(1 - q^{-nk})x^n}{n} W_k(x).$$

It is easy to show that $[W, P] = cW$ implies that $\exp(-P)W = \exp(c)\cdot W\exp(-P)$ as long as c commutes with both W and P. Therefore, we infer that

$$\begin{aligned}
& \exp^{-1} W_k(x) = \exp\left[\sum_{n=1}^{\infty}\frac{(1 - q^{-nk})x^n}{n}\right] W_k(x)\exp^{-1} \\
\Longrightarrow\quad & \exp^{-1} W_k(x) = \frac{1 - (x/q^k)}{1 - x}\cdot W_k(x)\exp^{-1} \\
\Longrightarrow\quad & \Phi_m \exp^{-1} W_k(x)\cdot(1 - x) = \Phi_m W_k(x)\exp^{-1}\cdot\left(1 - \frac{x}{q^k}\right).
\end{aligned}$$

With this in mind, (1-12) and the fact that $\Phi_m \exp^{-1} = A_m$ imply that

$$m^k W_k(x\gamma)\Phi_m \exp^{-1}\cdot\left(1 - \frac{x}{q^k}\right) = \Phi_m W_k(x)\exp^{-1}\cdot\left(1 - \frac{x}{q^k}\right)$$

Multiplying on the right with exp yields (1-17). $\square$

4 The Verma module

4.1 Let us now specialize to $S = \mathbb{A}^2$, and explain all the necessary modifications to the constructions in the present paper; we refer the reader to [14, Section 3] for details. From here on, let $\mathcal{M}$ be the moduli space parametrizing rank r torsion-free sheaves $\mathcal{F}$ on $\mathbb{P}^2$, together with a trivialization along a fixed line $\infty \subset \mathbb{P}^2$:

$$\mathcal{M} = \{\mathcal{F}, \mathcal{F}|_\infty \overset{\phi}{\cong} \mathcal{O}_\infty^r\}.$$

The c_1 of such sheaves is forced to be 0, but c_2 is free to vary over the nonnegative integers, and so the moduli space breaks up into connected components as before:

$$\mathcal{M} = \bigsqcup_{c=0}^{\infty} \mathcal{M}_c.$$

The space $\mathcal{M}$ is acted on by the torus $T = \mathbb{C}^* \times \mathbb{C}^* \times (\mathbb{C}^*)^r$, where the first two factors act by scaling $\mathbb{A}^2$, and the latter r factors act on the framing ϕ. Note that

$$K_0^T(\mathrm{pt}) = \mathbb{Z}[q_1^{\pm 1}, q_2^{\pm 1}, u_1^{\pm 1}, \dots, u_r^{\pm 1}],$$

where $q_1, q_2, u_1, \dots, u_r$ are the standard elementary characters of the torus T. We note that q_1 and q_2 are the equivariant weights of $\Omega^1_{\mathbb{A}^2}$, and the determinant of the universal sheaf $\mathcal{U}$ is the equivariant constant $u = u_1 \cdots u_r$. Consider the group

$$K_{\mathcal{M}} = \bigoplus_{c=0}^{\infty} K_0^T(\mathcal{M}_c) \otimes_{\mathbb{Z}[q_1^{\pm 1}, q_2^{\pm 1}, u_1^{\pm 1}, \dots, u_r^{\pm 1}]} \mathbb{Q}(q_1, q_2, u_1, \dots, u_r)$$

The main goal of loc. cit. was to define operators akin to (2-20), (2-21) and (2-23),

$$W_{n,k}, P_{\pm n'} \colon K_{\mathcal{M}} \to K_{\mathcal{M}} \tag{4-1}$$

for all $n \in \mathbb{Z}$ and $k, n' \in \mathbb{N}$, and then show that these operators satisfy the relations in the deformed W–algebra of type $\mathfrak{gl}_r$ (since $S = \mathbb{A}^2$, $K_{\mathcal{M}} \cong K_{\mathcal{M}\times S}$ naturally).

Definition 4.2 [14, Definition 2.28] Let $q_1, q_2, u_1, \dots, u_r$ be formal symbols. The universal Verma module $M_{u_1,\dots,u_r}$ is the $\mathbb{Q}(q_1, q_2, u_1, \dots, u_r)$–vector space with basis

$$W_{n_1,k_1} \dots W_{n_s,k_s} |\varnothing\rangle \tag{4-2}$$

as the pairs (n_i, k_i) range over $-\mathbb{N} \times \{1, \dots, r\}$ and are ordered in nondecreasing order of the slope n_i/k_i. We make $M_{u_1,\dots,u_r}$ into a deformed W–algebra module as follows.

The action of an arbitrary generator $W_{n,k}$ on the basis vector (4-2) is prescribed by the commutation relations (2-28), together with the relations

$$\begin{aligned} W_{n,k}|\varnothing\rangle &= 0 && \text{if } n>0 \text{ or } k>r, \\ W_{0,k}|\varnothing\rangle &= e_k(u_1,\dots,u_r)|\varnothing\rangle && \text{for all } k, \end{aligned}$$

where e_k denotes the k^{th} elementary symmetric polynomial.

Theorem 4.3 [14, Theorem 3.12] *We have an isomorphism of modules for the deformed W–algebra of type $\mathfrak{gl}_r$ (the action on the left-hand side is given by (4-1))*

$$K_{\mathcal{M}} \cong M_{u_1,\dots,u_r}, \tag{4-3}$$

induced by sending the K–theory class of the structure sheaf of $\mathcal{M}_0 \subset \mathcal{M}$ to $|\varnothing\rangle$.

4.4 The Ext (respectively vertex) operator A_m (respectively Φ_m) for $S=\mathbb{A}^2$ was studied in [14, Section 4], where we obtained an analogue of Theorem 1.7 in the case $k=1$ (some coefficients in the formulas of loc. cit. differ from those of the present paper, because their operator A_m differs from ours by an equivariant constant). However, having only proved the case $k=1$ in loc. cit. led to weaker formulas than (1-12). Thus, the present paper strengthens the results of loc. cit.; see Remark 4.8 therein. Specifically, Corollary 1.11 completely determines the operator Φ_m (hence also A_m) in the case $S=\mathbb{A}^2$, due to Theorems 4.3 and 4.5.

Theorem 4.5 *Given two Verma modules $M_{u_1,\dots,u_r}$ and $M_{u'_1,\dots,u'_r}$, there is a unique (up to constant multiple in $\mathbb{Q}(q_1,q_2,u_1,\dots,u_r,u'_1,\dots,u'_r)$) linear map*

$$\Phi_m\colon M_{u'_1,\dots,u'_r} \to M_{u_1,\dots,u_r}$$

satisfying (1-17) for all $k \geq 1$.

Proof The existence of such a linear map follows from the very fact that the operator (1-16) satisfies (1-17). To show uniqueness, it is enough to prove $\langle\varnothing|\Phi_m|\varnothing\rangle = 0$ implies $\Phi_m = 0$, for any operator that satisfies the following relations for all n, k:

$$\Phi_m W_{n,k} - \Phi_m W_{n+1,k}\cdot q^{-k} = W_{n,k}\Phi_m\cdot m^k\gamma^{-nk} - W_{n+1,k}\Phi_m\cdot\frac{m^k}{q^k}\gamma^{-(n+1)k}, \tag{4-4}$$

where m and γ are certain nonzero constants.

Claim 4.6 *For any parameters $u_1, \dots, u_r$, there exists a nondegenerate pairing*

$$M_{u_1,\dots,u_r} \otimes M_{u_1,\dots,u_r} \xrightarrow{\langle\cdot,\cdot\rangle} \mathbb{Q}(q_1, q_2, u_1, \dots, u_r)$$

such that the adjoint of $W_{n,k}$ is $W_{-n,k}$ for all $n \in \mathbb{Z}$ and $k \in \mathbb{N}$.

Proof Using Theorem 4.3, the required pairing is provided by the equivariant Euler characteristic pairing on $K_{\mathcal{M}}$ (renormalized as in [14, Section 3.14]). The operators $W_{n,k}$ and $W_{-n,k}$ are adjoint with respect to this pairing [14, formula (3.39)]. □

Let us now complete the proof of Theorem 4.5. Because Verma modules are generated by $W_{n,k}$ acting on $\varnothing$, then we must show that $\langle \varnothing | \Phi_m | \varnothing \rangle = 0$ implies

$$\langle \varnothing | W_{-n_s,k_s} \dots W_{-n_1,k_1} \Phi_m W_{n'_1,k'_1} \dots W_{n'_t,k'_t} | \varnothing \rangle = 0 \tag{4-5}$$

for all collections of indices $(n_i, k_i), (n'_i, k'_i) \in \mathbb{Z}_{\leq 0} \times \{1, \dots, r\}$, ordered by slope

$$\frac{n_1}{k_1} \leq \dots \leq \frac{n_s}{k_s} \quad \text{and} \quad \frac{n'_1}{k'_1} \leq \dots \leq \frac{n'_t}{k'_t}.$$

The matrix coefficient (4-5) is nonzero only if the n_i and the n'_j are all nonpositive, so we will prove formula (4-5) by induction on the nonpositive integer $\delta = \sum n_i + \sum n'_i$. We may assume that $n_s, n'_t < 0$ because $W_{0,k} | \varnothing \rangle$ is a multiple of $| \varnothing \rangle$ for any k. The base case $\delta = 0$ of the induction is simply the assumption $\langle \varnothing | \Phi_m | \varnothing \rangle = 0$. As for the induction step, let us iterate relation (4-4) to obtain

$$\Phi_m W_{n'_1,k'_1} \cdots W_{n'_t,k'_t} \in \operatorname{span} \begin{cases} \Phi_m W_{n'_1+\varepsilon_1,k'_1} \cdots W_{n'_t+\varepsilon_t,k'_t}, \\ W_{n'_1+\varepsilon'_1,k'_1} \cdots W_{n'_t+\varepsilon'_t,k'_t} \Phi_m, \end{cases}$$

where $\varepsilon_1, \dots, \varepsilon_t \in \{0, 1\}$ are not all 0, and $\varepsilon'_1, \dots, \varepsilon'_t \in \{0, 1\}$. That means that the left-hand side of (4-5) is a linear combination of

$$\langle \varnothing | W_{-n_s,k_s} \cdots W_{-n_1,k_1} \Phi_m W_{n'_1+\varepsilon_1,k'_1} \cdots W_{n'_t+\varepsilon_t,k'_t} | \varnothing \rangle,$$

which is 0 by the induction hypothesis, because the ε_i are not all 0, and

$$\langle \varnothing | W_{-n_s,k_s} \cdots W_{-n_1,k_1} W_{n'_1+\varepsilon'_1,k'_1} \cdots W_{n'_t+\varepsilon'_t,k'_t} \Phi_m | \varnothing \rangle. \tag{4-6}$$

The induction step will be complete once we show that (4-6) is 0. As a consequence of (2-28), the product of W's in (4-6) can be written as a linear combination of

$$W_{-n''_r,k''_r} \cdots W_{-n''_1,k''_1} \quad \text{with} \quad \frac{n''_1}{k''_1} \leq \dots \leq \frac{n''_r}{k''_r},$$

and $\sum n_i'' = \sum n_i - \sum n_i' - \sum \varepsilon_i'$ for degree reasons. If $n_r'' > 0$, then the product of W's above annihilates $\langle \varnothing |$. Thus, we may assume $n_r'' \leq 0$, in which case the fact that

$$\sum n_i'' = \sum n_i - \sum n_i' - \sum \varepsilon_i' > \sum n_i + \sum n_i'$$

(recall that $n_i' < 0$ by assumption, while $\varepsilon_i' \in \{0, 1\}$) means that we can apply the induction hypothesis to conclude that (4-6) is 0. □

We note that the identification of A_m (in the case $S = \mathbb{A}^2$) with a vertex operator was also achieved in [3], which computed relations (3-42) and (3-43) for $n = 1$ in the basis of fixed points. This uniquely determines the operator A_m due to certain features of the Ding–Iohara–Miki algebra, but does not directly establish the connection with the generating currents of the deformed W–algebra of $\mathfrak{gl}_r$. From a geometric point of view, this is because the Nakajima-type simple correspondences only describe the operators $L_{1,k}$ and $U_{1,k}$. As we have seen in Section 2.4, in order to define the operators $L_{n,k}$ and $U_{n,k}$ for all n (with the ultimate goal of defining the W–algebra generators $W_{n,k}$ in (2-20)), one needs to introduce the more complicated correspondences (2-11).

References

[1] **H Awata**, **H Kubo**, **S Odake**, **J Shiraishi**, *Quantum $\mathscr{W}_N$ algebras and Macdonald polynomials*, Comm. Math. Phys. 179 (1996) 401–416 MR Zbl

[2] **V Baranovsky**, *Moduli of sheaves on surfaces and action of the oscillator algebra*, J. Differential Geom. 55 (2000) 193–227 MR Zbl

[3] **J-E Bourgine**, **M Fukuda**, **Y Matsuo**, **H Zhang**, **R-D Zhu**, *Coherent states in quantum $\mathcal{W}_{1+\infty}$ algebra and qq-character for 5d super Yang–Mills*, Prog. Theor. Exp. Phys. (2016) 123B05, 41 MR Zbl

[4] **A Braverman**, **M Finkelberg**, **H Nakajima**, *Instanton moduli spaces and $\mathscr{W}$–algebras*, Astérisque 385, Soc. Math. France, Paris (2016) MR Zbl

[5] **U Bruzzo**, **M Pedrini**, **F Sala**, **R J Szabo**, *Framed sheaves on root stacks and supersymmetric gauge theories on ALE spaces*, Adv. Math. 288 (2016) 1175–1308 MR Zbl

[6] **E Carlsson**, **N Nekrasov**, **A Okounkov**, *Five-dimensional gauge theories and vertex operators*, Mosc. Math. J. 14 (2014) 39–61, 170 MR Zbl

[7] **E Carlsson**, **A Okounkov**, *Exts and vertex operators*, Duke Math. J. 161 (2012) 1797–1815 MR Zbl

[8] **B Feigin**, **E Frenkel**, *Quantum $\mathscr{W}$–algebras and elliptic algebras*, Comm. Math. Phys. 178 (1996) 653–678 MR Zbl

[9] **B Feigin**, **S Gukov**, VOA[M_4], J. Math. Phys. 61 (2020) 012302, 27 MR

[10] **L Göttsche**, **M Kool**, *Virtual refinements of the Vafa–Witten formula*, Comm. Math. Phys. 376 (2020) 1–49 MR Zbl

[11] **I Grojnowski**, *Instantons and affine algebras, I: The Hilbert scheme and vertex operators*, Math. Res. Lett. 3 (1996) 275–291 MR Zbl

[12] **D Maulik**, **A Okounkov**, *Quantum groups and quantum cohomology*, Astérisque 408, Soc. Math. France, Paris (2019) MR Zbl

[13] **H Nakajima**, *Heisenberg algebra and Hilbert schemes of points on projective surfaces*, Ann. of Math. 145 (1997) 379–388 MR Zbl

[14] **A Neguţ**, *The q–AGT-W relations via shuffle algebras*, Comm. Math. Phys. 358 (2018) 101–170 MR Zbl

[15] **A Neguţ**, *Shuffle algebras associated to surfaces*, Selecta Math. 25 (2019) art. id. 36 MR Zbl

[16] **A Neguţ**, *Hecke correspondences for smooth moduli spaces of sheaves*, Publ. Math. Inst. Hautes Études Sci. 135 (2022) 337–418 MR Zbl

[17] **A Neguţ**, *W–algebras associated to surfaces*, Proc. Lond. Math. Soc. 124 (2022) 601–679 MR Zbl

[18] **O Schiffmann**, **E Vasserot**, *Cherednik algebras, W–algebras and the equivariant cohomology of the moduli space of instantons on* $\mathbb{A}^2$, Publ. Math. Inst. Hautes Études Sci. 118 (2013) 213–342 MR Zbl

Department of Mathematics, MIT
Cambridge, MA, United States

Simion Stoilow Institute of Mathematics
Bucharest, Romania

andrei.negut@gmail.com

Proposed: Richard P Thomas — Received: 26 May 2020
Seconded: Lothar Göttsche, Mark Gross — Revised: 9 March 2022

Geometry & Topology 27:8 (2023) 3095–3181
DOI: 10.2140/gt.2023.27.3095
Published: 9 November 2023

Partially hyperbolic diffeomorphisms homotopic to the identity in dimension 3, II: Branching foliations

THOMAS BARTHELMÉ
SÉRGIO R FENLEY
STEVEN FRANKEL
RAFAEL POTRIE

We study 3–dimensional partially hyperbolic diffeomorphisms that are homotopic to the identity, focusing on the geometry and dynamics of Burago and Ivanov's center stable and center unstable *branching* foliations. This extends our previous study of the true foliations that appear in the dynamically coherent case. We complete the classification of such diffeomorphisms in Seifert fibered manifolds. In hyperbolic manifolds, we show that any such diffeomorphism is either dynamically coherent and has a power that is a discretized Anosov flow, or is of a new potential class called a *double translation*.

37D30, 57R30, 37C15, 57M50, 37D20

1 Introduction

A diffeomorphism f of a 3–manifold M is *partially hyperbolic* if it preserves a splitting of the tangent bundle TM into three 1–dimensional subbundles

$$TM = E^{\mathrm{s}} \oplus E^{\mathrm{c}} \oplus E^{\mathrm{u}},$$

where the *stable bundle* E^{s} is eventually contracted, the *unstable bundle* E^{u} is eventually expanded, and the *center bundle* E^{c} is distorted less than the stable and unstable bundles at each point. That is, for some $n > 0$ one has, at each $x \in M$,

$$\begin{gathered} \|Df^n|_{E^{\mathrm{s}}(x)}\| < 1, \\ \|Df^n|_{E^{\mathrm{u}}(x)}\| > 1, \\ \|Df^n|_{E^{\mathrm{s}}(x)}\| < \|Df^n|_{E^{\mathrm{c}}(x)}\| < \|Df^n|_{E^{\mathrm{u}}(x)}\|. \end{gathered}$$

From a geometric perspective, one can think of partial hyperbolicity as a generalization of the discrete behavior of an Anosov flow. On a 3–manifold M, such a flow Φ preserves a splitting of the unit tangent bundle TM into three 1–dimensional subbundles

$$TM = E^{\mathrm{s}} \oplus T\Phi \oplus E^{\mathrm{u}},$$

where E^{s} is eventually exponentially contracted, E^{u} is eventually exponentially expanded, and $T\Phi$ is the tangent direction to the flow. After flowing for a fixed time, an Anosov flow generates a partially hyperbolic diffeomorphism of a particularly simple type, where the stable and unstable bundles are contracted uniformly, and the center direction, which corresponds to $T\Phi$, is left undistorted. More generally, there are examples of partially hyperbolic diffeomorphisms of the form $f(x) = \Phi_{\tau(x)}(x)$, where Φ is a (topological) Anosov flow and $\tau\colon M \to \mathbb{R}_{>0}$ is a positive continuous function; the partially hyperbolic diffeomorphisms obtained in this way are called *discretized Anosov flows*.

A partially hyperbolic diffeomorphism is said to be *dynamically coherent* if there are invariant foliations tangent to the center stable and center unstable bundles $E^{\mathrm{c}} \oplus E^{\mathrm{s}}$ and $E^{\mathrm{c}} \oplus E^{\mathrm{u}}$. Discretized Anosov flows are dynamically coherent, since their center stable and center unstable bundles are uniquely integrable. On the other hand, we show in [3] that large classes of dynamically coherent partially hyperbolic diffeomorphisms must in fact be discretized Anosov flows:

Theorem 1.1 [3, Theorem A] *Let $f: M \to M$ be a dynamically coherent partially hyperbolic diffeomorphism on a closed Seifert fibered 3–manifold. If f is homotopic to the identity, then some iterate is a discretized Anosov flow.*

Theorem 1.2 [3, Theorem B] *Let $f: M \to M$ be a dynamically coherent partially hyperbolic diffeomorphism on a closed hyperbolic 3–manifold. Then some iterate is a discretized Anosov flow.*

The assumption of dynamical coherence is natural from a geometric perspective: the way that an Anosov flow distorts its weak stable and weak unstable foliations is often seen as the defining property of such a flow. In this light, the preceding results say that on certain classes of manifolds, any diffeomorphism with a geometric structure reminiscent to that of an Anosov flow must in fact come from one.

This assumption is much less satisfying from a dynamical perspective, however. Here the interest in partial hyperbolicity stems from its appearance as a generic consequence of dynamical conditions, such as stable ergodicity and robust transitivity (see Bonatti, Díaz and Viana [6]), and one is not provided with any invariant foliations. Although dynamical coherence was once generally expected, a number of recent results (see for example Barthelmé, Fenley, Frankel and Potrie [4], Bonatti, Gogolev, Hammerlindl and Potrie [7] and Rodriguez Hertz, Rodriguez Hertz and Ures [31]) have shattered that belief. For instance, in the unit tangent bundle of a hyperbolic surface, we proved in [4] that many partially hyperbolic diffeomorphisms are not dynamically coherent.

In our study of the dynamically coherent case in [3], the key to relating the inherently local property of partial hyperbolicity with the global structure of the ambient manifold lay in understanding the geometry and topology of the center stable and center unstable foliations, as well as their leafwise and transverse dynamics. The present article does away with the assumption of dynamical coherence. Instead of foliations we work with the center stable and center unstable "branching foliations" constructed by Burago and Ivanov [10] under certain orientability conditions. These are generalizations of foliations in which distinct leaves are allowed to merge together.

A large part of the present paper is concerned with carrying over our understanding of the geometry of foliations to branching foliations. We find that much of the familiar structure still holds in this more general context — sometimes by direct analogy, and sometimes with considerably more work. At the same time, there are important points at which branching foliations allow for more varied behavior than true foliations. A particularly important example of this appears in Figure 9, where the possibility of merging leaves thwarts one's ability to use the qualitative transverse and tangent behavior of a dynamical system to draw conclusions about its Lefschetz index. We hope that our work will entice those interested in the theory of foliations to consider the possible uses for branching foliations.

The following two theorems, which generalize the preceding theorems from [3], summarize the major consequences of the present article.

Theorem A *Let $f: M \to M$ be a partially hyperbolic diffeomorphism on a closed Seifert fibered 3–manifold. If f is homotopic to the identity, then it is dynamically coherent, and some iterate is a discretized Anosov flow.*

This is a stronger version of Theorem 1.1, without the a priori assumption of dynamical coherence. The following corresponds to Theorem 1.2.

Theorem B *Let $f: M \to M$ be a partially hyperbolic diffeomorphism on a closed hyperbolic 3–manifold. Then either*

(i) *f is dynamically coherent, some iterate is a discretized Anosov flow; or*

(ii) *f is not dynamically coherent, and after taking a finite cover[1] and iterate, it has center stable and center unstable branching foliations which are $\mathbb{R}$–covered and uniform, and a lift of f acts as a nontrivial translation on both of the corresponding leaf spaces.*

The existence or nonexistence of examples of type (ii) is one of the major questions coming out of this article. See Section 2.0.6.

Let us also mention a dynamical consequence of our analysis (Corollary 4.14).

Theorem 1.3 *Let $f: M \to M$ be a partially hyperbolic diffeomorphism of a closed 3–manifold M that is homotopic to the identity. If either M is hyperbolic or Seifert fibered, or the center stable or center unstable branching foliation is f–minimal, then f has no contractible periodic points* (see Definition 4.13).

[1] This is only needed to get the existence of f–invariant branching foliations.

Acknowledgements We thank C Bonatti, A Gogolev and A Hammerlindl for interesting discussions. We also thank the referee for their careful reading and thoughtful suggestions which led to significant improvements, especially in Section 6.

Barthelmé was partially supported by the NSERC, funding reference RGPIN-2017-04592. Fenley was partially supported by Simons Foundation grant 280429. Frankel was partially supported by National Science Foundation grant DMS-1611768. Potrie was partially supported by CSIC 618 and ANII–FCE–135352.

2 Outline and discussion

After recalling some definitions, we outline the more detailed results that lie behind our main theorems.

Let $f: M \to M$ be a partially hyperbolic diffeomorphism that is homotopic to the identity on a closed 3–manifold M.

Convention Throughout this paper, we will assume that the group $\pi_1(M)$ is not virtually solvable.

Although this assumption is not always necessary, it will simplify certain parts of the exposition. It does not result in loss of generality, since partially hyperbolic diffeomorphisms have been completely classified in manifolds with solvable or virtually solvable fundamental group; see Hammerlindl and Potrie [22; 23].

A foundational result of Burago and Ivanov (Theorem 3.6) implies that, after passing to an appropriate finite power and lift, we can assume that there is a pair of "branching foliations" $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ that are preserved by f and tangent to the center stable and center unstable bundles $E^{\mathrm{c}} \oplus E^{\mathrm{s}}$ and $E^{\mathrm{c}} \oplus E^{\mathrm{u}}$.

We outline the theory of these branching foliations in Section 3, and construct corresponding *leaf spaces* $\mathcal{L}^{\mathrm{cs}}$ and $\mathcal{L}^{\mathrm{cu}}$. Like the leaf spaces of true foliations, these are simply connected, possibly non-Hausdorff 1–manifolds that capture the transverse structure of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$, the lifts of $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ to the universal cover. This is where a large part of our work takes place, studying the dynamics of the following important class of lifts of f.

Definition 2.1 A lift of f to the universal cover is called *good* if it moves each point a uniformly bounded distance and commutes with every deck transformation.

Since f is homotopic to the identity, it has at least one good lift, obtained by lifting such a homotopy.

Remark 2.2 The diffeomorphisms we consider are in fact *isotopic* to identity: indeed, all the manifolds that appear in this article are irreducible and covered by $\mathbb{R}^3$. Then, the works of many authors — Waldhausen [35] for Haken manifolds, Boileau and Otal [5] for Seifert manifolds and Gabai, Meyerhoff and Thurston [21] for hyperbolic manifolds — give that homotopy implies isotopy. We will however not use this fact in the sequel, as the existence of a good lift is all that we use.

2.0.1 Dynamics on leaf spaces In Section 4, we study the way that good lifts of f permute the leaves of the lifted center stable and center unstable branching foliations, and the implications for the structure of their leaf spaces. This extends [3, Section 3].

The picture is particularly simple when $\mathcal{W}^{cs}$ is f–minimal, which means that the only closed, nonempty, f–invariant set which is a union of leaves is M itself. If $\mathcal{W}^{cs}$ is f–minimal, then:

($\star$)
- Each good lift $\widetilde{f}$ fixes either every leaf or no leaf of $\widetilde{\mathcal{W}}^{cs}$.
- If some good lift $\widetilde{f}$ fixes no leaf, then $\mathcal{W}^{cs}$ is $\mathbb{R}$–covered and uniform, and $\widetilde{f}$ acts as a translation its leaf space.

The same holds for $\widetilde{\mathcal{W}}^{cu}$. In particular, if both $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$ are f–minimal, then one of the following holds for each good lift $\widetilde{f}$ of f:

(1) **Double invariance** $\widetilde{f}$ fixes every leaf of both $\widetilde{\mathcal{W}}^{cs}$ and $\widetilde{\mathcal{W}}^{cu}$.

(2) **Mixed behavior** $\widetilde{f}$ fixes every leaf of either $\widetilde{\mathcal{W}}^{cs}$ or $\widetilde{\mathcal{W}}^{cu}$, and acts as a translation on the leaf space of the other.

(3) **Double translation** $\widetilde{f}$ acts as a translation on the leaf spaces of both $\widetilde{\mathcal{W}}^{cs}$ and $\widetilde{\mathcal{W}}^{cu}$.

This trichotomy applies whenever f is transitive or volume-preserving, where the associated branching foliations are always f–minimal [8].

When f is a discretized Anosov flow, there is a natural homotopy from the identity to f that moves points along the orbits of the underlying flow. The good lift $\widetilde{f}$ that comes from lifting this homotopy fixes every center leaf. In order to show that a given partially hyperbolic diffeomorphism is a discretized Anosov flow, we will need to find a good lift with this property. Here, one takes the center leaves to be the components of intersections between center stable and center unstable leaves. In particular, we will need find a good lift with doubly invariant behavior.

2.0.2 Center dynamics in fixed leaves In Section 5, we study the dynamics of the center foliation within center stable and center unstable leaves. We obtain the following crucial tool (see Definition 5.1 and Proposition 5.2):

($\star\star$) Suppose that $\mathcal{W}^{cs}$ is f–minimal, and that some good lift $\tilde{f}$ fixes every center stable leaf but no center leaf in $\widetilde{M}$. Then every f–periodic center leaf in M is coarsely contracted.

If one replaces $\mathcal{W}^{cs}$ with $\mathcal{W}^{cu}$ then one concludes that any f–periodic center leaf in M is coarsely expanded. This is widely applicable since one can find a periodic center leaf on any center stable or center unstable leaf with nontrivial fundamental group (Proposition 5.6).

Remark 2.3 In the dynamically coherent case, ($\star\star$) leads to a contradiction that yields a *fixed* center leaf [3, Proposition 4.4]. In Section 9 we show that this holds as well under the assumption of *absolute* partial hyperbolicity.

2.0.3 Minimality in hyperbolic and Seifert fibered manifolds In Section 6, we show the following, which means that the preceding trichotomy holds whenever the ambient manifold is hyperbolic or Seifert fibered.

If M is hyperbolic or Seifert fibered, then:

($\star'$)
- Each good lift $\tilde{f}$ fixes either every leaf or no leaf of $\widetilde{\mathcal{W}}^{cs}$.
- If some good lift $\tilde{f}$ fixes every leaf, then $\mathcal{W}^{cs}$ is f–minimal.
- If some good lift $\tilde{f}$ fixes no leaf, then $\mathcal{W}^{cs}$ is $\mathbb{R}$–covered and uniform, and $\tilde{f}$ acts as a translation on its leaf space.

2.0.4 Double invariance implies dynamical coherence In Section 7 we prove the following criterion for when a partially hyperbolic diffeomorphism is a discretized Anosov flow:

Theorem 2.4 *Let $f\colon M \to M$ be a partially hyperbolic diffeomorphism that is homotopic to the identity. If f admits f–minimal center stable and center unstable branching foliations, and some good lift $\tilde{f}$ has doubly invariant behavior, then f is a discretized Anosov flow.*

The key is to show that such an f is dynamically coherent. Then [3, Theorem 6.1] implies that it is a discretized Anosov flow.

Until this point we have always assumed that the bundles E^{s}, E^{c} and E^{u} have orientations that are preserved by f so that we can use the result of Burago and Ivanov to find center stable and center unstable branching foliations. In Section 7.3, we show that if a lift of an iterate of f is dynamically coherent and has a good lift $\widetilde{g}$ with doubly invariant behavior, then f is dynamically coherent. This is why Theorems A and B(i) do not need the orientability conditions.

2.0.5 Seifert fibered and hyperbolic manifolds We rule out mixed behavior in Seifert fibered manifolds in Section 8, and in hyperbolic manifolds in Sections 11–12. Together with Theorem 2.4, this yields the following:

Theorem 2.5 *Let $f : M \to M$ be a partially hyperbolic diffeomorphism homotopic to the identity on a closed hyperbolic or Seifert fibered 3–manifold. Assume that there are center stable and center unstable branching foliations. Then each good lift of f either*

(i) *fixes every leaf of both $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$, or*

(ii) *acts as a translation on both leaf spaces.*

If there is a good lift of type (i), *then f is a discretized Anosov flow.*

As was already pointed out in [3, Remark 7.3], there are examples in Seifert fibered manifolds where every good lift acts as a double translation. However, we show in Section 8 that one can always find a finite power of such diffeomorphisms with a good lift that has doubly invariant behavior. Together with the results of Section 7 this implies Theorem A.

Since every diffeomorphism of a hyperbolic 3–manifold has an iterate homotopic to the identity, one also deduces Theorem B.

Remark 2.6 An analogue of Theorem 2.5 holds under the assumption of f–minimality together with *absolute* partial hyperbolicity. See Section 9.

We believe that Theorem 2.5 should hold, using the same strategy as here, under the assumption of f–minimality together with the existence of an atoroidal piece in the JSJ decomposition of M. We have not pursued this here as it would require proving results similar to [33; 11; 17] in this setting.

2.0.6 Double translations This leaves open one major question:

Question Is there a partially hyperbolic diffeomorphism on a closed hyperbolic 3–manifold whose good lifts act as double translations?

As noted above, there are such examples on Seifert fibered manifolds, but by Theorem A these are all dynamically coherent and have iterates that are discretized Anosov flows.

The dynamics of a double translation on a hyperbolic manifold would have to be coarsely comparable to that of a pseudo-Anosov flow; see Section 11. The closest analogues from this perspective are the non-dynamically-coherent examples on Seifert manifolds, constructed in [7], which act as pseudo-Anosov maps on the base.

2.1 Remarks and references

There are three major areas in which the general case differs significantly from the dynamically coherent case:

(1) Unlike the dynamically coherent case (see condition ($\star\star$) in [3, Section 2]), there may be annular center stable leaves which do not contain a closed center leaf.

(2) In hyperbolic manifolds, we cannot rule out the possibility of double translations from the general version of the existence of cores that "shadow" the periodic orbits of the transverse pseudo-Anosov flow; see condition ($\star\star\star$) in [3, Section 2].

(3) In hyperbolic and Seifert manifolds, it is more difficult to eliminate the hypothesis of f–minimality. See Section 6.

We refer to [15; 24; 30] for surveys on the problem of classification of partially hyperbolic diffeomorphisms in dimension 3. There is earlier work towards classification that does not assume dynamical coherence, but these articles tend to have two simplifying characteristics: they work with manifolds on which taut foliations are well understood and amenable to classification, and on which known partially hyperbolic models are available for comparison. Typically, dynamical coherence is established under the assumption of nonexistence of invariant tori by using the fact that coarse dynamics separates leaves of the branching foliations. Neither of these features hold for the classes of manifolds considered in this article, and dynamical incoherence may appear in several different ways.

For instance, we obtain dynamical coherence in Section 7 when the lift of the partially hyperbolic diffeomorphism fixes each leaf of the lifted branching foliations. We also learn more about the structure of the branching foliations in the non-dynamically-coherent case, leading, in particular, to case (ii) of Theorem B. This structure also allows us to better understand the dynamical properties of the system, even when the manifold is not hyperbolic or Seifert fibered, as can be seen in Theorem 1.3.

More generally, the framework that we develop for the study of non-dynamically-coherent partially hyperbolic diffeomorphism is useful outside of the homotopy class of the identity.

Below are several tools developed in this article that we wish to emphasize:

(1) In Sections 3 and 4, we develop some of the basic theory necessary for the topological study of branching foliations and the diffeomorphisms that preserve them, including the structure of their leaf spaces.

(2) In Section 5.1 we introduce the notion of coarsely contracting and coarsely repelling periodic rays. This should be useful for the study of all partially hyperbolic diffeomorphisms in 3–manifolds, ie including those not homotopic to the identity.

(3) In Section 6 we study the way that certain special lifts of a partially hyperbolic diffeomorphism act within a fixed center stable leaf, and find conditions that guarantee the nonexistence of fixed points. This involves understanding the behavior of strong stable manifolds through fixed points under iteration, which may find applications in other contexts.

(4) In Section 7 we prove uniqueness of (branching) foliations under certain conditions. This is a key to finding results that do not require taking finite lifts and finite powers. As such, it may also be relevant for the study of topological obstructions for partially hyperbolic diffeomorphisms — note that the topological obstructions for the existence of Anosov flows can depend on taking finite lifts; see eg [12].

There is other work that shows the uniqueness of branching foliations, but always in a setting where there is an understood model partially hyperbolic diffeomorphism for comparison.

(5) In Sections 11 and 12 we develop some tools to analyze the transverse geometry of branching foliations. This combines ideas from the theory of Lefschetz index, hyperbolic geometry, and the notion of coarsely expanding and contracting rays in item (2).

The tools in (5) are used in [4] to prove that a large class of partially hyperbolic diffeomorphisms in Seifert manifolds are dynamically incoherent. In addition, (2) and (5) are used in [18] to obtain fine dynamical consequences of partial hyperbolicity in 3–manifolds.

3 Branching foliations and leaf spaces

In this section we review the existence of center stable and center unstable branching foliations, and construct corresponding leaf spaces that capture their transverse topology. We will also construct a "center foliation" and leaf space.

Definition 3.1 A *branching foliation* of a 3–manifold M is a collection $\mathcal{F}$ of C^1–immersed surfaces, called *leaves*, each complete in its induced metric, such that

(i) each $x \in M$ is contained in at least one leaf,

(ii) no leaf crosses itself,

(iii) different leaves do not cross each other, and

(iv) if L_n are leaves, and $x_n \in L_n$ converges to a point $x \in M$, then some subsequence of the L_n converges to a leaf L with $x \in L$.[2]

Here, "crossing" is meant in a topological sense; see [10] or [24].

Remark 3.2 In this context, "branching" refers to the fact that leaves may merge. This should not be confused with the typical use of "branching" in the theory of codimension-1 foliations, where it refers to non-Hausdorff behavior in the leaf space.

Since a branching foliation has C^1 leaves that do not cross, it has a well-defined tangent distribution.

As with foliations, there is a sense in which branching foliations are "locally product (branched) foliated": around each point one can find a neighborhood U with a smooth product structure $U \simeq \mathbb{D}^2 \times [0,1]$ such that each leaf of $\mathcal{F}$ that intersects U does so in a collection of discs that are transverse to the $[0,1]$–fibration and meet every $[0,1]$–fiber. This follows readily from the fact that branching foliations are tangent to C^1 distributions.

On a compact manifold there is a uniform scale ϵ_0, called the *local product structure size*, such that every open set of diameter less than ϵ_0 is contained in a product chart as above.

[2]Here, convergence should be understood in the pointed compact–open topology, ie given a compact set K in L containing x, there is a sequence of compact subsets K_n of L_n containing x_n such that K_n converges to K in the Hausdorff topology.

Definition 3.3 A branching foliation $\mathcal{F}$ is *well-approximated by foliations* if there is, for a set of $\epsilon > 0$ accumulating on 0, a family of foliations $\{\mathcal{F}_\epsilon\}$ with C^1 leaves, and a family of continuous maps $\{h_\epsilon \colon M \to M\}$, which have the following properties (with respect to some fixed Riemannian metric):

(v) The angles between leaves of $\mathcal{F}$ and $\mathcal{F}_\epsilon$ are less than ϵ.

(vi) The C^0–distance between h_ϵ and the identity is less than ϵ.

(vii) On each leaf of $\mathcal{F}_\epsilon$, the map h_ϵ restricts to a local diffeomorphism to a leaf of $\mathcal{F}$.

(viii) For each leaf L of $\mathcal{F}$ there is a leaf L_ϵ of $\mathcal{F}_\epsilon$ with $h_\epsilon(L_\epsilon) = L$.

Remark 3.4 While the maps h_ϵ restrict to local diffeomorphisms on leaves, they will fail to be global diffeomorphisms on leaves of $\mathcal{F}_\epsilon$ that map to self-merging leaves of $\mathcal{F}$. In addition, the h_ϵ will not be local diffeomorphisms on M unless $\mathcal{F}$ is actually a true foliation.

Definition 3.5 A partially hyperbolic diffeomorphism $f \colon M \to M$ is said to be *orientable* if the bundles E^{s}, E^{u} and E^{c} admit orientations that are preserved by f.

The following is the foundational existence result of Burago and Ivanov:

Theorem 3.6 (Burago and Ivanov [10]) *Let f be an orientable partially hyperbolic diffeomorphism of a 3–manifold M. Then there are f–invariant branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ tangent to $E^{\mathrm{c}} \oplus E^{\mathrm{s}}$ and $E^{\mathrm{c}} \oplus E^{\mathrm{u}}$ that are well-approximated by foliations.*

Here, a branching foliation is said to be *f–invariant* if the image of any leaf under f is again a leaf.

Note that there is no a priori uniqueness for the center stable and center unstable branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ related to a partially hyperbolic diffeomorphism. Nevertheless, we will typically fix *some* pair of such branching foliations and call them "the" branching foliations for our diffeomorphism. In addition, we will fix families of approximating foliations $\mathcal{W}^{\mathrm{cs}}_\epsilon$ and $\mathcal{W}^{\mathrm{cu}}_\epsilon$, with associated maps denoted by h^{cs}_ϵ and h^{cu}_ϵ.

On the other hand, since the stable bundle E^{s} *is* uniquely integrable, a stable leaf s that intersects a center stable leaf L must be contained entirely in L. Consequently, the intersection of any two center stable leaves is saturated by stable leaves.

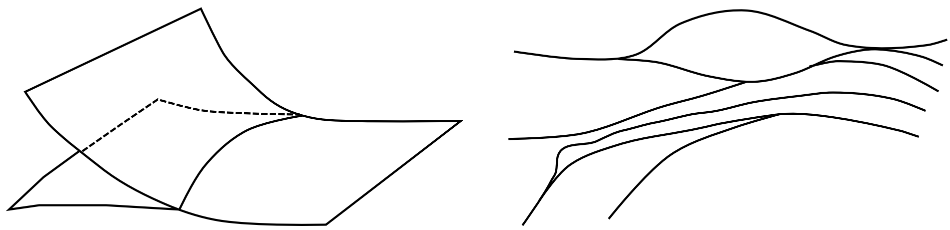

Figure 1: The branching of center and center stable leaves. Left: two center stable leaves sharing a region. Right: distinct center leaves inside a center stable leaf.

Once we have fixed "the" center stable and center unstable branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$, the corresponding lifted foliations on $\widetilde{M}$ will be denoted by $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$. We may then define center leaves as follows:

Definition 3.7 A *center leaf* of a partially hyperbolic diffeomorphism is the projection to M of a connected component of the intersection between a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$.

Although the collection of center leaves is not a foliation, it is a kind of codimension-2 branching foliation. We will abuse terminology and call the collection of center leaves the *center foliation*.

Remark 3.8 Each center leaf is tangent to the central direction E^{c}, but a complete curve that is tangent to the central direction may not be a center leaf. Indeed, even when the diffeomorphism is dynamically coherent, the central direction may not be uniquely integrable. See [31] for an example.

3.1 Tautness

In this article, the approximating foliations $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$ and $\mathcal{W}^{\mathrm{cu}}_{\epsilon}$ have no compact leaves.

Indeed, suppose that one has a compact leaf $L \in \mathcal{W}^{\mathrm{cs}}_{\epsilon}$. Then $K := h^{\mathrm{cs}}_{\epsilon}(L)$ is a compact leaf of $\mathcal{W}^{\mathrm{cs}}$. Since the stable bundle E^{s} is uniquely integrable, this compact surface has a foliation without compact leaves, so it is a torus. According to [27, Theorem 1.4], there are only a few classes of manifolds that admit partially hyperbolic diffeomorphisms with tori tangent to $E^{\mathrm{s}} \oplus E^{\mathrm{c}}$, all mapping tori of $\mathbb{T}^2$.

Since we assume that $\pi_1(M)$ is not virtually solvable, it follows that the approximating foliations have no compact leaves, which implies that they are taut.

3.2 Center stable and center unstable leaf spaces

Given a foliation $\mathcal{F}$ on a manifold M, the set of leaves of the lifted foliation $\widetilde{\mathcal{F}}$ on $\widetilde{M}$ has a natural topology — the quotient obtained from $\widetilde{M}$ by collapsing each leaf to a point — and the resulting space is called the *leaf space* of M.

In this section we will define a notion of leaf space for our branching foliations, where it would not make sense to take the quotient topology. We will see, in fact, that the leaf spaces of our branching foliations are homeomorphic to those of the approximating foliations for small enough ϵ.

Much of this section would apply to any codimension-1 branching foliation, of any dimension, as long as the leaves in the universal cover are properly embedded $\mathbb{R}^{n-1}$'s in $\mathbb{R}^n$. For convenience, however, we will mostly restrict attention to the branching foliations that we are interested in. This allows for some shortcuts. For example, in Proposition 3.16 we use the approximating foliations and maps to see that the leaf space is a 1–manifold as desired, though this could also be done directly.

3.2.1 Complementary regions and sides Since M is not finitely covered by $S^2 \times S^1$ (as $\pi_1(M)$ is not virtually solvable), and our branching foliations are well-approximated by taut foliations, it follows that the universal cover is homeomorphic to $\mathbb{R}^3$, and the lifted leaves are properly embedded planes [14].

The *complementary regions* of a leaf L are the two connected components of $\widetilde{M} \setminus L$. For each complementary region U of a leaf L, the closure $\overline{U} = U \cup L$ is called a *side* of L.

A coorientation of the branching foliation (which may be thought of as a coorientation of its tangent distribution) determines, for each leaf L, a *positive* and a *negative* complementary region, which we denote by $L^{\oplus}$ and $L^{\ominus}$. The corresponding sides are denoted by $L^+ = L^{\oplus} \cup L$ and $L^- = L^{\ominus} \cup L$. We will fix such a coorientation throughout.

3.2.2 Leaf spaces Let us now construct the *center stable leaf space* $\mathcal{L}^{\mathrm{cs}}$. This is the set of leaves of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ with the topology defined below. The *center unstable leaf space* $\mathcal{L}^{\mathrm{cu}}$ is constructed similarly.

In the case of a true codimension-1 foliation, each transverse arc in the universal cover maps homeomorphically to an arc in the leaf space. We will use a similar idea for branching foliations, and use transverse arcs to construct the topology. In a true foliation each point in a transverse arc intersects a single leaf; for our branching foliations we need to "blow up" at some points, using the following definition:

Definition 3.9 Given $x \in \widetilde{M}$, let $\mathcal{L}^{\mathrm{cs}}(x) \subset \mathcal{L}^{\mathrm{cs}}$ denote the set of leaves that contain x.

Given distinct leaves $L \neq E$ in $\mathcal{L}^{\mathrm{cs}}(x)$, we will write $L <_x E$ whenever $L^+ \supset E$.

Claim 3.10 *For each $x \in \widetilde{M}$, $<_x$ defines a linear order, with respect to which $\mathcal{L}^{\mathrm{cs}}(x)$ is order-isomorphic to a closed interval (possibly a single point).*

Proof Assume that $\mathcal{L}^{\mathrm{cs}}(x)$ is not a singleton.

That $<_x$ defines a linear order on $\mathcal{L}^{\mathrm{cs}}(x)$ follows from the fact that leaves do not cross (property (iii) of Definition 3.1). From property (iv), it follows that this order is complete.

To see that $\mathcal{L}^{\mathrm{cs}}(x)$ is order-isomorphic to a closed interval, it suffices to check that there are no gaps in the order. That is, given $L, E \in \mathcal{L}^{\mathrm{cs}}(x)$ such that $L <_x E$, we must find some $L' \in \mathcal{L}^{\mathrm{cs}}(x)$ with $L <_x L' <_x E$.

Given such L and E, let y be a boundary point of the connected component of $L \cap E$ that contains x. Consider a neighborhood B of y with diameter less than ϵ_0, the local product structure size of $\mathcal{W}^{\mathrm{cs}}$. Since $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ is product branched foliated in B, each leaf that intersects $B \cap (L^+ \cap E^-)$ must intersect y, and since leaves do not cross, any such leaf must intersect x. Any such leaf L' will have $L <_x L' <_x E$. □

Combined with the linear ordering of points in a transversal, this gives a linear ordering on the set of leaves that intersect a transversal:

Definition 3.11 Given a transverse arc τ, let $\mathcal{L}^{\mathrm{cs}}(\tau) \subset \mathcal{L}^{\mathrm{cs}}$ denote the set of leaves that intersect τ.

Orient τ so that it agrees with the coorientation on $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Given distinct leaves $K \neq L$ in $\mathcal{L}^{\mathrm{cs}}(\tau)$, we will write $K <_\tau L$ whenever either

- $K \cap \tau$ lies forward of $L \cap \tau$ with respect to the orientation on τ, or
- K and L intersect τ at the same point x and $K <_x L$.

The following properties of these orderings may be found in [10, Section 7].

Claim 3.12 (1) *For each* **open** *transverse arc* τ, $<_\tau$ *is a linear order, with respect to which* $\mathcal{L}^{\mathrm{cs}}(\tau)$ *is order-isomorphic to an open interval.*

(2) *If* σ *and* τ *are open transverse arcs, then* $<_\sigma$ *and* $<_\tau$ *define the same linear order on* $\mathcal{L}^{\mathrm{cs}}(\sigma) \cap \mathcal{L}^{\mathrm{cs}}(\tau)$, *which is order-isomorphic to an open interval (possibly empty).*

Definition 3.13 (topology of $\mathcal{L}^{\mathrm{cs}}$) The *center stable leaf space* is $\mathcal{L}^{\mathrm{cs}}$, with the topology $\mathcal{T}$ generated by all open intervals in $\mathcal{L}^{\mathrm{cs}}(\tau) \subset \mathcal{L}^{\mathrm{cs}}$, over all open transverse arcs τ.

From Claim 3.12(2), it suffices to take any collection of open transverse arcs that intersect every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Since M is compact, one can take a finite collection of open transverse arcs in M and consider all of their lifts to $\widetilde{M}$. This implies in particular that $\mathcal{L}^{\mathrm{cs}}$ is second-countable.

Proposition 3.14 *The center stable leaf space* $\mathcal{L}^{\mathrm{cs}}$ *is a simply connected, possibly non-Hausdorff* 1*–manifold.*

The same applies to $\mathcal{L}^{\mathrm{cu}}$. This is not difficult to prove directly, and it applies more generally to any codimension-1 branching foliation of a closed n–manifold, as long as the lifted foliation is by properly embedded $\mathbb{R}^{n-1}$'s in $\widetilde{M} \simeq \mathbb{R}^n$. In the present case, it follows as well from Proposition 3.16 below.

3.2.3 Leaf spaces and approximating foliations Let $\mathcal{L}^{\mathrm{cs}}_\epsilon$ and $\mathcal{L}^{\mathrm{cu}}_\epsilon$ denote the leaf spaces of the approximating foliations $\mathcal{W}^{\mathrm{cs}}_\epsilon$ and $\mathcal{W}^{\mathrm{cu}}_\epsilon$. The maps h^{cs}_ϵ and h^{cu}_ϵ induce functions

$$g_{\epsilon,s} \colon \mathcal{L}^{\mathrm{cs}}_\epsilon \to \mathcal{L}^{\mathrm{cs}} \quad \text{and} \quad g_{\epsilon,u} \colon \mathcal{L}^{\mathrm{cu}}_\epsilon \to \mathcal{L}^{\mathrm{cu}}$$

between the corresponding leaf spaces, which are surjective whenever ϵ is sufficiently small; cf Definition 3.3.

Since $\mathcal{W}^{\mathrm{cs}}_\epsilon$ is a true foliation, its leaf space $\mathcal{L}^{\mathrm{cs}}_\epsilon$ is a simply connected, possibly non-Hausdorff 1–manifold; cf [3, Appendix B].

Remark 3.15 It is possible to modify the proof of [10, Theorem 7.2], where the foliations $\mathcal{W}^{\mathrm{cs}}_\epsilon$ and maps h^{cs}_ϵ are constructed, so that the $g_{\epsilon,s}$ are injective in addition to surjective. With this in hand, one could define the topology on $\mathcal{L}^{\mathrm{cs}}$ to be the one induced by this bijection.

Instead of redoing the entire proof of [10, Theorem 7.2], we will use a simpler fact that can easily be extracted from that proof: the maps h^{cs}_ϵ are "monotone" in the sense that they preserve the natural linear order on plaques in local charts.

Proposition 3.16 *When ϵ is sufficiently small,*

(1) *the preimage of each point in $\mathcal{L}^{cs}$ under $g_{\epsilon,s}$ is a closed interval,*

(2) *$g_{\epsilon,s}\colon \mathcal{L}^{cs}_\epsilon \to \mathcal{L}^{cs}$ is continuous, and*

(3) *the topology $\mathcal{T}$ on $\mathcal{L}^{cs}$ is equivalent to the quotient topology $\mathcal{T}_\epsilon$ induced by $g_{\epsilon,s}$.*

The same applies for the center unstable foliations.

Proof Let ϵ_0 be the local product sizes of $\mathcal{W}^{cs}$, and let $\epsilon < \epsilon_0/2$. Let $\mathcal{T}_\epsilon$ be the quotient topology induced by $g_{\epsilon,s}$ on $\mathcal{L}^{cs}$.

(1) Let $I \subset \mathcal{L}^{cs}_\epsilon$ be the preimage of a leaf $L \in \mathcal{L}^{cs}$, and suppose that I contains two leaves $\widehat{L}_1$ and $\widehat{L}_2$. We want to show that $\widetilde{h}^{cs}_\epsilon$ takes every leaf between $\widehat{L}_1$ and $\widehat{L}_2$ to L.

From property (vi) of Definition 3.3, the Hausdorff distance between $\widehat{L}_1$ and $\widehat{L}_2$ is less than 2ϵ. Since 2ϵ was chosen to be less than the local product structure size, it follows that the region between $\widehat{L}_1$ and $\widehat{L}_2$ has leaf space which is a closed interval. By the local monotonicity of $\widetilde{h}^{cs}_\epsilon$, it follows that $g_{\epsilon,s}$ maps the entire region between $\widehat{L}_1$ and $\widehat{L}_2$ to L. This implies that the preimage of L is an interval, which is closed because $\widetilde{h}^{cs}_\epsilon$ is continuous.

(2) Let $U \subset \mathcal{L}^{cs}$ be open. Around each point in U one can find an open interval $J \subset U$ that is the set of leaves intersecting a small open transversal β. We want to show that $g^{-1}_{\epsilon,s}(J)$ is open in $\mathcal{L}^{cs}_\epsilon$.

Let $\widehat{L}_1$ be a leaf in $g^{-1}_{\epsilon,s}(J)$. Then $\widehat{L}_1$ intersects β (or a slightly bigger transversal), so all the leaves of $\widetilde{\mathcal{W}}^{cs}_\epsilon$ close enough to $\widehat{L}_1 \cap \beta$ intersect β. Thus an open neighborhood of $\widehat{L}_1$ is contained in $g^{-1}_{\epsilon,s}(J)$, and $g_{\epsilon,s}$ is continuous.

(3) From (2) it follows that $\mathcal{T} \subset \mathcal{T}_\epsilon$. Let us prove the other inclusion.

Suppose $W \in \mathcal{L}^{cs}$ is an open set in $\mathcal{T}_\epsilon$, and let $y \in W$. Then $U = (g_{\epsilon,s})^{-1}(W)$ is an open set containing the closed interval $I = (g_{\epsilon,s})^{-1}(y)$. Let L and E be the boundary leaves of I. Then one can find half-open intervals $I_L, I_E \subset U$ such that $I_L \cap I = L$ and $I_E \cap I = E$. Then $I_L \cup I \cup I_E$ projects to a set in $\mathcal{L}^{cs}$ which contains an open interval around y in $\mathcal{L}^{cs}$. Since this applies for every $y \in W$ it follows that W is open in $\mathcal{T}$. □

This suffices to show that $\mathcal{L}^{\mathrm{cs}}$ is a 1–manifold. It is possible to modify $g_{\epsilon,s}\colon \mathcal{L}^{\mathrm{cs}}_{\epsilon} \to \mathcal{L}^{\mathrm{cs}}$ to be a homeomorphism when ϵ is sufficiently small, but we will not need this fact.

In the sequel, we fix ϵ small enough that the previous proposition applies for both the center stable and center unstable foliations.

3.3 Center "foliations"

3.3.1 The center foliation within a center stable/unstable leaf Fix a center stable leaf L of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. We will describe the topology of the center leaf space, $\mathcal{L}^{\mathrm{c}}_{L}$, restricted to L. The center leaf within a center unstable leaf is defined in the same manner.

Remark 3.17 Recall from Definition 3.7 that a center leaf in $\widetilde{M}$ is defined as a connected component of the intersection between a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$. Now, the following situation may arise (see Figure 2): two leaves U_1, U_2 of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$ and a leaf L of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ such that the triple intersection $U_1 \cap L \cap U_2$ contains a connected component of c_1 of $U_1 \cap L$ as well as a connected component c_2 of $U_2 \cap L$. That is, the center leaves c_1 and c_2 represent the same set in $\widetilde{M}$. In this case, we also consider c_1 and c_2 as the *same* leaf of the center foliation in L.

Definition 3.18 (topology $\mathcal{A}$ in $\mathcal{L}^{\mathrm{c}}_{L}$) Consider a countable set of open transversals τ_i which are perpendicular to the center bundle in L, and whose union intersects every center leaf in L. Put the order topology on the set I_i of center leaves intersecting τ_i. This induces the topology $\mathcal{A}$ in $\mathcal{L}^{\mathrm{c}}_{L}$.

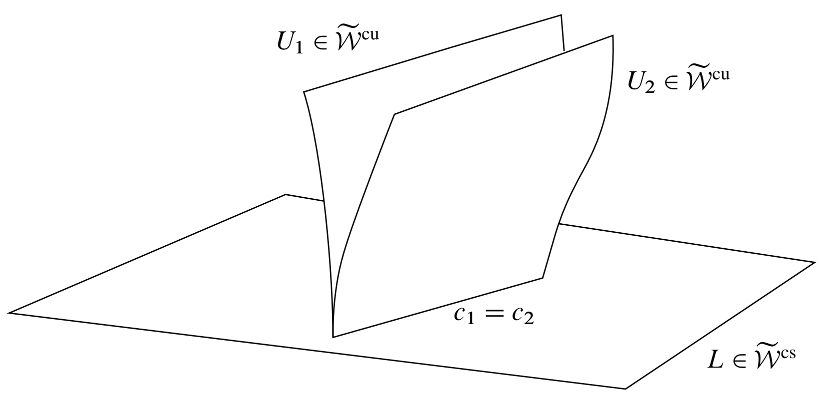

Figure 2: Different center unstable leaves may intersect a given center stable leaf in the same center leaf.

Let L be a fixed leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. We again fix an $\epsilon > 0$ and consider the approximating foliation $\widetilde{\mathcal{W}}^{\mathrm{cu}}_\epsilon$. Since $\widetilde{\mathcal{W}}^{\mathrm{cu}}$ is transverse to L, so is $\widetilde{\mathcal{W}}^{\mathrm{cu}}_\epsilon$ (for ϵ small enough). Thus, $\widetilde{\mathcal{W}}^{\mathrm{cu}}_\epsilon$ induces a 1–dimensional (nonbranching) foliation $\mathcal{F}_\epsilon$ on L, and hence its leaf space $\mathcal{L}^{\mathrm{c}}_{L,\epsilon}$ is a 1–dimensional, not necessarily Hausdorff, simply connected manifold.

The behavior described in Remark 3.17 above leads to the following issue: the unique center leaf $c_1 = c_2$ is approximated by two distinct leaves of $\mathcal{F}_\epsilon$. Thus, the leaf space $\mathcal{L}^{\mathrm{c}}_L$ of the center foliation on L is not in bijection with $\mathcal{L}^{\mathrm{c}}_\epsilon$. However, we still have a surjective, but not necessarily injective, projection $\mathrm{pr}_\epsilon \colon \mathcal{L}^{\mathrm{c}}_{L,\epsilon} \to \mathcal{L}^{\mathrm{c}}_L$ as in the previous subsection. Let $\mathcal{A}_\epsilon$ be the quotient topology from the map pr_ϵ.

Just as in Proposition 3.16 one can prove the following:

Lemma 3.19 *The set of center leaves in L through a point x is a closed interval. Let c_0 be a center leaf in L. Let $I = \mathrm{pr}^{-1}(c_0) \subset \mathcal{L}^{\mathrm{c}}_\epsilon$. The set I is a closed interval. If $\epsilon < \epsilon_0$, then the topologies $\mathcal{A}$ and $\mathcal{A}_\epsilon$ are the same.*

3.3.2 Center foliation in $\widetilde{M}$ Finally, we have to put a topology on the leaf space $\mathcal{L}^{\mathrm{c}}$ of the center foliation in $\widetilde{M}$.

Pick an $0 < \epsilon < \epsilon_0$ so that $\widetilde{\mathcal{W}}^{\mathrm{cs}}_\epsilon$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}_\epsilon$ are transverse to each other. Call $\mathcal{F}_\epsilon$ the 1–dimensional foliation obtained as the intersection of $\widetilde{\mathcal{W}}^{\mathrm{cs}}_\epsilon$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}_\epsilon$. The leaf space $\mathcal{L}^{\mathrm{c}}_\epsilon$ of $\mathcal{F}_\epsilon$ is now a simply connected, possibly non-Hausdorff, 2–dimensional manifold. But as before, there is only a surjective, and not injective, projection $g_\epsilon \colon \mathcal{L}^{\mathrm{c}}_\epsilon \to \mathcal{L}^{\mathrm{c}}$.

The map g_ϵ is defined in the following way: if $\overline{c}$ is a leaf of $\mathcal{F}_\epsilon$, then it is the intersection of a leaf $\overline{U}$ of $\widetilde{\mathcal{W}}^{\mathrm{cu}}_\epsilon$ and a leaf $\overline{S}$ of $\widetilde{\mathcal{W}}^{\mathrm{cs}}_\epsilon$. There exists a unique connected component c of $g_{\epsilon,u}(\overline{U}) \cap g_{\epsilon,s}(\overline{S})$ that is at distance less than 2ϵ from $\overline{c}$. We define $g_\epsilon(\overline{c}) = c$.

Once again, the topology $\mathcal{B}_\epsilon$ we put on $\mathcal{L}^{\mathrm{c}}$ is obtained by identifying elements of $\mathcal{L}^{\mathrm{c}}_\epsilon$ that project to the same element of $\mathcal{L}^{\mathrm{c}}$ and taking the quotient topology.

As done previously in Sections 3.2.2 and 3.3.1, in order to prove that the topology that we put on $\mathcal{L}^{\mathrm{c}}$ makes it a simply connected (not necessarily Hausdorff) 2–manifold, it is enough to show that the preimages of points by g_ϵ are closed, simply connected sets contained in a local chart of $\mathcal{L}^{\mathrm{c}}_\epsilon$. In order to do that, first notice that $\mathcal{L}^{\mathrm{c}}_\epsilon$ is locally homeomorphic to $\mathcal{L}^{\mathrm{cs}}_\epsilon \times \mathcal{L}^{\mathrm{cu}}_\epsilon$. Indeed, any $\overline{c}_0 \in \mathcal{L}^{\mathrm{c}}_\epsilon$ is a connected component of $\overline{U}_0 \cap \overline{S}_0$, with $\overline{U}_0 \in \mathcal{L}^{\mathrm{cu}}_\epsilon$ and $\overline{S}_0 \in \mathcal{L}^{\mathrm{cs}}_\epsilon$. Now, if V_u is a small enough open interval in $\mathcal{L}^{\mathrm{cu}}_\epsilon$ and V_s is a small enough open interval in $\mathcal{L}^{\mathrm{cs}}_\epsilon$, then for any $\overline{U} \in V_u$ and $\overline{S} \in V_s$, the intersection $\bar{U} \cap \bar{S}$ contains a unique connected component close to c_0. Using this local homeomorphism, the following lemma will imply that the topology $\mathcal{L}^{\mathrm{c}}$ is as we claimed.

Lemma 3.20 *Let c_0 be in $\mathcal{L}^{\mathrm{c}}$. The set $R = g_\epsilon^{-1}(c_0)$ is homeomorphic to a closed rectangle in $\mathcal{L}^{\mathrm{cs}}_\epsilon \times \mathcal{L}^{\mathrm{cu}}_\epsilon$.*

Proof Let $\bar{c}_1, \bar{c}_2 \in R$. Let $\bar{U}_1$ be the leaf in $\mathcal{L}^{\mathrm{cu}}_\epsilon$ containing $\bar{c}_1$ and let $\bar{S}_2$ be the leaf in $\mathcal{L}^{\mathrm{cs}}_\epsilon$ containing $\bar{c}_2$. Let $U_1 = g_{\epsilon,u}(\bar{U}_1)$ and $S_2 = g_{\epsilon,s}(\bar{S}_2)$. Since $\bar{c}_1, \bar{c}_2 \in R$, the center leaf c_0 is a connected component of $U_1 \cap S_2$. Thus $\bar{U}_1$ and $\bar{S}_2$ must intersect and the intersection contains a unique connected component $\bar{c}_3$ at distance at most 2ϵ from c_0.

Now, the proof of Lemma 3.19 shows that $\bar{c}_1$ and $\bar{c}_3$ are two ends of an interval in the leaf space of $\mathcal{F}_\epsilon$ restricted to $\bar{U}_1$ that is entirely contained in R, and similarly for $\bar{c}_2$ and $\bar{c}_3$ considered as elements of the leaf space of $\mathcal{F}_\epsilon$ restricted to $\bar{S}_2$. In turn, the arguments of the proof of Lemma 3.19 imply that the set R projects to a closed interval in both $\mathcal{L}^{\mathrm{cs}}_\epsilon$ and $\mathcal{L}^{\mathrm{cu}}_\epsilon$, ie it is a closed rectangle in $\mathcal{L}^{\mathrm{cs}}_\epsilon \times \mathcal{L}^{\mathrm{cu}}_\epsilon$. □

Just as in the previous two sections we can also put a topology $\mathcal{B}$ on $\mathcal{L}^{\mathrm{c}}$ directly as follows:

Definition 3.21 (topology $\mathcal{B}$ on $\mathcal{L}^{\mathrm{c}}$) In M pick a collection of very small open rectangles R_i which are almost perpendicular to the center bundle, and with boundary two arcs in leaves of $\mathcal{L}^{\mathrm{cs}}$ and two arcs in leaves of $\mathcal{L}^{\mathrm{cu}}$. Consider all lifts R of these to $\widetilde{M}$. The set of center leaves intersecting R is naturally bijective to an open rectangle and we give it the topology making this a local homeomorphism. The topology $\mathcal{B}$ is generated by these rectangles.

First we justify why the set of center leaves through R is naturally an open rectangle. Let L_1, L_2 be the center stable leaves containing the two arcs in the boundary of R, and U_1, U_2 be the corresponding center unstable leaves. The set of center stable leaves between L_1, L_2 (not including L_1, L_2) is naturally order-isomorphic to an open interval. This was proved in Section 3.2.2. The same holds for the center unstable foliation. The product is an open rectangle. The set of center leaves intersecting R is a quotient of this. The sets which are quotiented to a point are compact subrectangles. The proof is the same as the previous lemma. Hence the quotient is naturally a rectangle. In addition, if a collection of center leaves intersects two such rectangles R and R', then the identifications in R also produce the same identifications in R', and the order of the center stable and center unstable foliations in the subsets are the same whether in R or R'. Hence in the identification, the topologies agree.

Just as in the previous sections one can prove:

Lemma 3.22 *For $\epsilon < \epsilon_0$, the topologies $\mathcal{B}$ and $\mathcal{B}_\epsilon$ are the same.*

The main property is to prove is exactly that of Lemma 3.20. The rest follows just as in the previous subsections.

3.4 From foliations to branching foliations

Using the leaf space, one can carry over a number of concepts from foliations to branching foliations.

3.4.1 Uniform and $\mathbb{R}$–covered branching foliations A branching foliation is said to be $\mathbb{R}$–*covered* if its leaf space is homeomorphic to $\mathbb{R}$. It is *uniform* if every two leaves in the universal cover are a finite Hausdorff distance apart.

By Proposition 3.16 a branching foliation is uniform or $\mathbb{R}$–covered if and only if its approximating foliations are, for ϵ sufficiently small.

3.4.2 Saturations and minimality A foliation preserved by a homeomorphism f is said to be f–minimal if the only closed, saturated, f–invariant sets are the empty set and the whole manifold. We will define f–minimality identically for branching foliations, but we must be careful about what we mean by "saturated":

Definition 3.23 A set $C \subset M$ is $\mathcal{W}^{cs}$–*saturated* if, for every $x \in C$, there is a leaf of $\mathcal{W}^{cs}$ that contains x and is contained in C.

A *saturation* of a saturated set $C \subset M$ is a collection of leaves $X \subset \mathcal{W}^{cs}$ whose union is C.

Note that this is much weaker than asking that *every* leaf intersecting C be contained in C. In particular, our notion of saturation has the peculiar property that the complement of a saturated set need not be saturated; see Figure 3.

In addition, a saturated set may have different saturations. However, a saturated set always has a unique *maximal saturation*, consisting of all leaves that are contained in it.

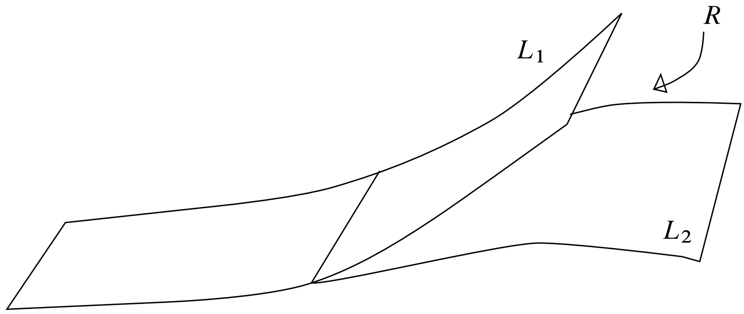

Figure 3: L_1 and L_2 are two leaves in C, but the region R is not in C. Then, in parts of R, all the center stable leaves intersect the branch locus between L_1 and L_2, so have parts in C and parts not in C (and therefore $M \setminus C$ is not saturated by center stable leaves).

Definition 3.24 We say that $\mathcal{W}^{\mathrm{cs}}$ is *f–minimal* if the only closed, $\mathcal{W}^{\mathrm{cs}}$–saturated, and f–invariant subsets of M are $\varnothing$ and M.

We emphasize that "closed" is meant as a subset of M, not $\mathcal{L}^{\mathrm{cs}}$.

Saturated sets and saturations are defined similarly in the universal cover. Here, a saturation can be naturally thought of as a subset of the leaf space $\mathcal{L}^{\mathrm{cs}}$. However, the topology of a saturated set in $\widetilde{M}$ does not necessarily agree with the topology of a saturation in $\mathcal{L}^{\mathrm{cs}}$:

Remark 3.25 Let $C \subset \widetilde{M}$ be $\widetilde{\mathcal{W}}^{\mathrm{cs}}$–saturated. It is possible for C to be closed in $\widetilde{M}$, but have a saturation $\mathcal{C} \subset \mathcal{L}^{\mathrm{cs}}$ that is not closed in $\mathcal{L}^{\mathrm{cs}}$. However, it is easy to see that C is closed in $\widetilde{M}$ if and only if its *maximal* saturation is closed in $\mathcal{L}^{\mathrm{cs}}$.

It is true but less immediate that the only saturation of $\widetilde{M}$ that is closed in $\mathcal{L}^{\mathrm{cs}}$ is all of $\mathcal{L}^{\mathrm{cs}}$ (Lemma B.1).

3.4.3 Perfect fits

The notion of "perfect fits" from the theory of codimension-1 foliations [3, Section 4.1] applies to branching foliations once it is modified appropriately.

We will need the 2–dimensional version of this concept, in Section 5, to understand the center and stable foliations within a center stable leaf. Given a center stable leaf L, let $\mathcal{C}_L$ and $\mathcal{S}_L$ be the center and stable foliations within L, and let $\mathcal{L}^{\mathrm{c}}_L$ and $\mathcal{L}^{\mathrm{s}}_L$ be the corresponding leaf spaces.

Definition 3.26 A leaf $c \in \mathcal{C}_L$ and a leaf $s \in \mathcal{S}_L$ make a *$\mathcal{CS}$–perfect fit* if they do not intersect, but there is a local transversal τ to $\mathcal{C}_L$ through c such that every leaf in $\mathcal{C}_L(\tau)$ that lies sufficiently close to one side of c (in the linear order $<_\tau$) intersects s.

They make an *$\mathcal{SC}$–perfect fit* if there is a local transversal τ' to $\mathcal{S}_L$ through s such that every leaf in $\mathcal{S}_L(\tau)$ that lies sufficiently close to one side of s intersects c.

We say that c and s make a *perfect fit* if they make both a $\mathcal{CS}$– and $\mathcal{SC}$–perfect fit.

Remark 3.27 When defining $\mathcal{CS}$–perfect fits it is important to use the linear order $<_\tau$ on $\mathcal{C}_L(\tau)$, defined in Section 3.2.2, since there may be center leaves on the same side of c as s that merge with c.

Since $\mathcal{S}_L$ is a true foliation, the linear order $<_{\tau'}$ on $\mathcal{S}_L(\tau')$ comes directly from the transversal τ', so the notion of a $\mathcal{SC}$–perfect fit is exactly as in [3, Section 4.1].

One may equivalently define $\mathcal{CS}$–perfect fits as follows. Given a stable leaf s in L, let $I_s \subset \mathcal{L}^{\mathrm{c}}_L$ be the set of center leaves that intersect s. Then c and s make a $\mathcal{CS}$–perfect fit if and only if $c \in \partial I_s$.

Lemma 3.28 *Let c and s be center and stable leaves in a center stable leaf L that make a $\mathcal{CS}$–perfect fit. Then there is a stable leaf s' such that c and s' make a perfect fit.*

The symmetric statement holds for $\mathcal{SC}$–perfect fits.

Proof This is [3, Lemma 4.2], whose proof remains valid with the obvious modifications. □

4 Branching foliations and good lifts

Fix a closed 3–manifold M whose fundamental group is not virtually solvable, a partially hyperbolic diffeomorphism $f\colon M \to M$ homotopic to the identity, and a good lift $\widetilde{f}$. We will assume that f is orientable (Definition 3.5) so that we have center stable and center unstable branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ which are well-approximated by taut foliations (Theorem 3.6). This can be achieved by taking an iterate of f and lifting to a finite cover of M — we will deal with the effects of replacing f and M in Section 7.

In this section we will study the way that a good lift $\widetilde{f}$ acts on the lifted branching foliations $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$ in the universal cover $\widetilde{M}$.

4.1 Translation-like behavior

In this section, we will see that the action of $\widetilde{f}$ on the center stable leaf space must look locally like a translation. Identical statements hold for the center unstable foliation.

Remark 4.1 In fact, the results in this subsection are not really particular to partially hyperbolic diffeomorphisms. They apply to any diffeomorphism that is homotopic to the identity and that preserves a branching foliation well-approximated by taut foliations. In addition, in this subsection we also do not need to assume that $\pi_1(M)$ is virtually solvable.

The key to this section is the following fact:

Lemma 4.2 (big half-space lemma) *Let L be a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. For any $R > 0$, there exists a ball of radius R contained in each complementary region of L.*

Proof This lemma holds for true foliations — see [3, Lemma 3.3] — so it suffices to consider a leaf corresponding to L in the approximating foliation $\widetilde{\mathcal{W}}^{\mathrm{cs}}_{\epsilon}$ for ϵ sufficiently small. $\square$

Remark 4.3 The tautness of the foliation is essential for this result to hold. The branching foliations in the non-dynamically-coherent example of [31], for instance, do not satisfy that lemma.

Definition 4.4 (regions between leaves) Let $K, L \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$ be distinct leaves. In the leaf space, $\mathcal{L}^{\mathrm{cs}} \setminus \{K, L\}$ consists of three open connected components. Only one of these components accumulates on both K and L: we call this the *open $\mathcal{L}^{\mathrm{cs}}$–region between K and L*. Its closure in $\mathcal{L}^{\mathrm{cs}}$, which is obtained by adjoining K and L, is called the *closed $\mathcal{L}^{\mathrm{cs}}$–region between K and L*.

Remark 4.5 The subset of $\widetilde{M}$ that corresponds to the open $\mathcal{L}^{\mathrm{cs}}$–region between two leaves may not be open. However, the subset of $\widetilde{M}$ that corresponds to the closed $\mathcal{L}^{\mathrm{cs}}$–region between two leaves is closed. It is also connected, but its interior may not be. See Figure 4.

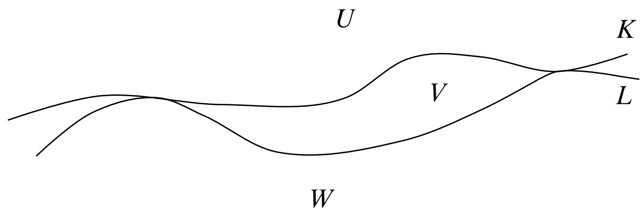

Figure 4: The interior of the closed region between leaves may not be connected.

The following is the equivalent of [3, Proposition 3.5]. The same proof applies if one considers complementary regions and regions between leaves as subsets of $\widetilde{M}$ and $\mathcal{L}^{cs}$, as appropriate.

Proposition 4.6 *If $L \in \widetilde{\mathcal{W}}^{cs}$ is not fixed by a good lift $\widetilde{f}$, then*

(1) *the closed $\mathcal{L}^{cs}$–region between L and $\widetilde{f}(L)$ is an interval,*
(2) *$\widetilde{f}$ takes each coorientation at L to the corresponding coorientation at $\widetilde{f}(L)$, and*
(3) *the subset of $\widetilde{M}$ corresponding to the closed $\mathcal{L}^{cs}$–region between L and $\widetilde{f}(L)$ is contained in the closed $2R$–neighborhood of L, where*

$$R = \max_{y \in \widetilde{M}} d(y, \widetilde{f}(y)).$$

Remark 4.7 In the above proposition, we may a priori have that L and $\widetilde{f}(L)$ merge.

Using Proposition 4.6 we therefore also obtain the equivalent of [3, Proposition 3.7].

Proposition 4.8 *The set $\Lambda \subset \mathcal{L}^{cs}$ of leaves that are fixed by $\widetilde{f}$ is closed and $\pi_1(M)$–invariant. Each connected component I of $\mathcal{L}^{cs} \setminus \Lambda$ is acted on by $\widetilde{f}$ as a translation, and every pair of leaves in I are a finite Hausdorff distance apart.*

In the above proposition, one has to be mindful again that "open" and "closed" refer to the topology on the leaf space $\mathcal{L}^{cs}$, and not the topology on $\widetilde{M}$.

When $\mathcal{W}^{cs}$ is f–minimal (Definition 3.24), we deduce the following dichotomy from Proposition 4.8:

Corollary 4.9 *If $\mathcal{W}^{cs}$ is f–minimal, then either*

(1) *$\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{cs}$, or*
(2) *$\mathcal{W}^{cs}$ is $\mathbb{R}$–covered and uniform, and $\widetilde{f}$ acts as a translation on the leaf space $\mathcal{L}^{cs}$.*

Proof Although the proof is conceptually identical to that of the corresponding result in the dynamically coherent case [3, Corollary 3.10], we will redo it since the distinction between the topology in $\mathcal{L}^{cs}$ and $\widetilde{M}$ becomes important.

Let Λ be the set of leaves that are fixed by $\tilde{f}$. Since $\tilde{f}$ commutes with deck transformations, each deck transformation preserves Λ. In particular, if I is a component of $\mathcal{L}^{cs} \setminus \Lambda$ and $g \in \pi_1(M)$, one has either $g(I) = I$ or $g(I) \cap I = \varnothing$.

So Λ is invariant under $\tilde{f}$ and deck transformations, saturated by $\widetilde{\mathcal{W}}^{cs}$, and closed in $\mathcal{L}^{cs}$ (by Proposition 4.8).

Let $\widetilde{B} \subset \widetilde{M}$ be the union of the points in all leaves in $\widetilde{\Lambda}$, and let $B = \pi(\widetilde{B}) \subset M$. Since Λ is closed in $\mathcal{L}^{cs}$, $\widetilde{B}$ is closed in $\widetilde{M}$, and B is closed in M. In addition, B is f–invariant. Since $\mathcal{W}^{cs}$ is f–minimal, B is either $\varnothing$ or M.

If B is empty then Λ is empty, and Proposition 4.8 implies that we are in case (2).

If $B = M$ then $\widetilde{B} = \widetilde{M}$, and we have to prove that $\Lambda = \mathcal{L}^{cs}$. This follows from the more general Lemma B.1, but it also has a more direct proof, as follows.

Suppose $\Lambda \neq \mathcal{L}^{cs}$. Let I be a connected component of $\mathcal{L}^{cs} \setminus \Lambda$. Let J be the set of points of $\widetilde{M}$ contained in a leaf in I. The set I is open (in $\mathcal{L}^{cs}$) and $\tilde{f}$ translates leaves in I. It follows that the interior in $\widetilde{M}$ of J is nonempty. These points in the interior of J are not contained in $\widetilde{B}$. This contradicts $\widetilde{B} = \widetilde{M}$. So $\Lambda = \mathcal{L}^{cs}$, and we are in case (1). □

This immediately implies the trichotomy in Section 2.0.1.

4.2 Ruling out fixed points

Let us now find conditions under which we show that our good lift $\tilde{f}$ has no fixed points in $\widetilde{M}$. We will use the following lemma.

Lemma 4.10 *Let $L \in \widetilde{\mathcal{W}}^{cs}$ be a center stable leaf that is fixed by $\tilde{f}$. Suppose that for every $y \in L$ one can find a leaf $L' \in \widetilde{\mathcal{W}}^{cs}$ that is fixed by $\tilde{f}$ and intersects the unstable leaf through y in a point other than y. Then no nontrivial power of $\tilde{f}$ fixes a point in L.*

Proof Suppose that $\tilde{f}^n$ fixes a point $x \in L$ for some $n \neq 0$. One can assume after possibly switching signs that $n > 0$. Then expansion of the unstable leaf u through x implies that no leaf L' that intersects u at a point other than x can be fixed. □

Compare this with the simpler statement [3, Lemma 3.13] in the dynamically coherent setting, where it suffices to assume L is not isolated in the set of fixed leaves.

Corollary 4.11 *If $\widetilde{f}$ fixes every center stable leaf, then it has no fixed or periodic points in $\widetilde{M}$.*

This follows immediately from the lemma. We will now exclude the existence of fixed or periodic points under the assumption of f–minimality.

Theorem 4.12 *If $\mathcal{W}^{\mathrm{cs}}$ or $\mathcal{W}^{\mathrm{cu}}$ is f–minimal, then $\widetilde{f}$ does not have any fixed or periodic points in $\widetilde{M}$.*

Proof Assume without loss of generality that $\mathcal{W}^{\mathrm{cs}}$ is f–minimal. By the dichotomy in Corollary 4.9, $\widetilde{f}$ either fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, or acts as a translation on $\mathcal{L}^{\mathrm{cs}}$.

If $\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, the result follows from Lemma 4.10. If $\widetilde{f}$ acts as a translation on $\mathcal{L}^{\mathrm{cs}}$, then for any leaf L of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ one has $\widetilde{f}^i(L) \cap L = \varnothing$ for $|i|$ sufficiently large. $\square$

A noteworthy consequence is the nonexistence of "contractible periodic points" under the assumption of f–minimality.

Definition 4.13 Let g be a homeomorphism of a manifold homotopic to the identity. A point p is a *contractible periodic point* if $g^n(p) = p$ for some $n \neq 0$ and there is a homotopy $H \colon M \times [0,1] \to M$ from the identity to g such that the concatenation of the paths $H(p,\cdot), H(g(p),\cdot), \dots, H(g^{n-1}(p),\cdot)$ is homotopically trivial.

Notice that if p is a contractible periodic point of g of period n then there exists a good lift $\widetilde{g}$ of g and a lift $\widetilde{p}$ of p such that $\widetilde{g}^n(\widetilde{p}) = \widetilde{p}$. Thus, Theorem 4.12 immediately yields:

Corollary 4.14 *If f admits a f–minimal branching center stable or center unstable foliation, then f has no contractible periodic points.*

This completes the proof of Theorem 1.3 in the f–minimal case. The hyperbolic and Seifert fibered cases follow from Proposition 6.1.

4.3 Fundamental groups of leaves

The leaves of $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ are immersed surfaces which may not be injectively immersed. In the universal cover, however, the leaves of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$ are properly embedded planes; cf Section 3.2.

It follows that there may be a closed loop in a leaf with a corresponding element of $\pi_1(M)$ that fixes no lift of that leaf in the universal cover. These elements are not useful for our purposes, so we will remove them by convention:

Convention When working with a fixed lift L of a leaf C of $\mathcal{W}^{\mathrm{cs}}$ or $\mathcal{W}^{\mathrm{cu}}$, we will say that an element $\gamma \in \pi_1(M)$ is in the fundamental group of C if it stabilizes L.

There is another way of seeing this notion of fundamental group arise: recall from Theorem 3.6 that the branching foliations are approximated by true foliations $\mathcal{W}^{\mathrm{cu}}_\epsilon$ and $\mathcal{W}^{\mathrm{cs}}_\epsilon$ and that there exists maps h^{cs}_ϵ and h^{cu}_ϵ mapping leaves of $\mathcal{W}^{\mathrm{cs}}_\epsilon$ (or $\mathcal{W}^{\mathrm{cu}}_\epsilon$) to those of $\mathcal{W}^{\mathrm{cs}}$ (or $\mathcal{W}^{\mathrm{cu}}$). Then, a loop is in the fundamental group of a leaf C of $\mathcal{W}^{\mathrm{cs}}$ if and only if it is freely homotopic to a loop in a corresponding leaf C_ϵ of $\mathcal{W}^{\mathrm{cs}}_\epsilon$ for every ϵ small enough. Notice that if there are several leaves that project to C, in the universal cover, take a lift L and it follows from Proposition 3.16 that the set of leaves that projects to L is an interval in the leaf space of $\widetilde{\mathcal{W}}^{\mathrm{cs}}_\epsilon$. It follows that h^{cs}_ϵ lifts to an equivariant (with respect to the defined fundamental group of C) diffeomorphism from the boundary leaves of the closed interval to L. We call such a leaf L_ϵ and write $C_\epsilon = \pi(L_\epsilon)$.

In other words, for us, the fundamental group of C based at y will be exactly $(h^{\mathrm{cs}}_\epsilon)_*(\pi_1(C_\epsilon, y_0))$, where $h^{\mathrm{cs}}_\epsilon(y_0) = y$.

In particular, since $\mathcal{W}^{\mathrm{cs}}_\epsilon$ and $\mathcal{W}^{\mathrm{cu}}_\epsilon$ are taut foliations without Reeb components, each leaf is π_1–injective in M. Thus, this second interpretation helps explain our convention: the closed loops in a leaf of $\mathcal{W}^{\mathrm{cs}}$ are either in the fundamental group as we defined it, or they are due to a self-intersection. In that case, they are not an essential feature of the leaf, as they stopped being closed when pulled-back to the approximating leaf.

Following our convention, we will then say that a leaf C of the branching foliation is a plane, a cylinder or a Möbius band if its corresponding approximated leaf C_ϵ is, respectively, a plane, a cylinder or a Möbius band for any small enough ϵ.

Using these conventions, Proposition 3.14 of [3] holds for the leaves of the branching foliations whenever $\widetilde{f}$ has no fixed points in the leaf; cf Lemma 4.10. For ease of reference, we restate it here.

Proposition 4.15 *Assume that $\widetilde{f}$ fixes a leaf L of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Then $C = \pi(L)$ has cyclic fundamental group* (*thus it is either a plane, an annulus or a Möbius band*), *or L has a point fixed by $\widetilde{f}$.*

Remark 4.16 Similarly, because of possible self-intersections, we need to be careful with how to define the path metric on a leaf of $\mathcal{W}^{\mathrm{cs}}$ or $\mathcal{W}^{\mathrm{cu}}$.

If C is a leaf of, say, $\mathcal{W}^{\mathrm{cs}}$, we define a *path* on C as a continuous curve η that is the projection of a continuous curve $\widetilde{\eta}$ in a lift L of C to $\widetilde{M}$. We then define the path metric on C as usual, but considering only the paths as defined before.

Notice that not every continuous curve η on C is a path in the above sense, as there might not exists any lift of η that stays on only one lift of C.

Still, the analogue of [3, Lemma 3.11] holds:

Lemma 4.17 *If $\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$* (*resp. $\widetilde{\mathcal{W}}^{\mathrm{cu}}$*) *then there is a $K > 0$ such that for every $L \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$* (*resp. $L \in \widetilde{\mathcal{W}}^{\mathrm{cu}}$*), *we have that $d_L(x, \widetilde{f}(x)) < K$.*

4.4 Gromov hyperbolicity of leaves

We now prove a version of [3, Lemma 3.20] in the non-dynamically-coherent setting.

Lemma 4.18 *If $\mathcal{W}^{\mathrm{cs}}$ is f–minimal, and $\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Then each leaf of $\mathcal{W}^{\mathrm{cs}}$ is Gromov hyperbolic.*

Proof The foliation $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$ is taut. Thus, Candel's theorem [13] asserts that either all the leaves of $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$ are Gromov hyperbolic or there is a holonomy-invariant transverse measure (of zero Euler characteristic).

Assume for a contradiction that μ is a holonomy-invariant transverse measure. Since $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$ is not f–invariant, we have to adjust the proof given in [3]. The transverse measure μ lifts to a measure $\widetilde{\mu}$ transverse to $\widetilde{\mathcal{W}}^{\mathrm{cs}}_{\epsilon}$. Thus, $\widetilde{\mu}$ defines a measure on $\mathcal{L}^{\mathrm{cs}}_{\epsilon}$, the leaf space of $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$.

Let $g_{\epsilon,s} : \mathcal{L}^{\mathrm{cs}}_{\epsilon} \to \mathcal{L}^{\mathrm{cs}}$ be the canonical projection between the leaf spaces of $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$ and $\mathcal{W}^{\mathrm{cs}}$; see Section 3.2.2. Let $\widetilde{\nu} := (g_{\epsilon,s})_* \widetilde{\mu}$ be the corresponding measure on $\mathcal{L}^{\mathrm{cs}}$. Now $\widetilde{\nu}$ is $\widetilde{f}$–invariant since $\widetilde{f}$ is the identity on $\mathcal{L}^{\mathrm{cs}}$, and it is also $\pi_1(M)$–invariant as $\widetilde{\mu}$ is. The support of $\widetilde{\nu}$ in $\mathcal{L}^{\mathrm{cs}}$ is a closed set Z in $\mathcal{L}^{\mathrm{cs}}$ that is $\widetilde{f}$–invariant and $\pi_1(M)$–invariant.

The measure $\widetilde{\nu}$ on $\mathcal{L}^{\mathrm{cs}}$ can also be considered as a measure on the set of transversals to $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ in $\widetilde{M}$: For any transversal τ to $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ in $\widetilde{M}$, we define $\widetilde{\nu}(\tau)$ as the $\widetilde{\nu}$–measure of the set of leaves in $\mathcal{L}^{\mathrm{cs}}$ that intersects τ. Notice that the measure of a point in $\widetilde{M}$ (which can be thought of as a degenerate transversal) can be positive if the image of that point in $\mathcal{L}^{\mathrm{cs}}$ is an interval.

Note also that we refrained from calling $\widetilde{\nu}$ a transverse measure to $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ because it is by no means holonomy-invariant. In fact holonomy itself is not well-defined for a branching foliation. Still, $\widetilde{\nu}$ satisfies the property that if τ_1, τ_2 are transversals and every leaf intersecting τ_1 also intersects τ_2, then $\widetilde{\nu}(\tau_1) \le \widetilde{\nu}(\tau_2)$.

Projecting down to M, the measure $\widetilde{\nu}$ induces a measure ν on the set of transversals to $\mathcal{W}^{\mathrm{cs}}$ on M.

Let τ be any unstable segment in M. Since $\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, the measure of $f^i(\tau)$ $(= \nu(f^i(\tau)))$ is equal to $\nu(\tau)$ for any integer i. We can choose i very big and negative so that the length of $f^i(\tau)$ is extremely small. Therefore it is contained in a small foliated box of $\mathcal{W}^{\mathrm{cs}}$, which is the projection of a compact foliated box of $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$. It follows that $\nu(\tau)$ is uniformly bounded. In particular, this implies that the ν–measure of any unstable leaf in M is bounded above. In turn, it implies that for any $j > 0$ (assumed big enough), there is an unstable segment u_j of length $> j$ which has $\nu(u_j)$ measure $< 1/j$. Taking the midpoint of these segments and a converging subsequence, we obtain a full unstable leaf, call it ζ, so that ζ has $\nu(\zeta) = 0$ (since $\nu(\zeta) < 1/j$ for all big enough j).

Let Y be the union of the leaves of $\mathcal{W}^{\mathrm{cs}}$ that do not intersect ζ or any of its iterates by f. Then Y is a closed subset of M and clearly f–invariant. Let L be a leaf in $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ which is in Z, the support of $\widetilde{\nu}$. Then by definition of support of $\widetilde{\nu}$, it follows that $\pi(L)$ cannot intersect ζ or any of its iterates by f. Hence $\pi(L)$ is in Y. In particular, Y is not empty. This contradicts the fact that $\mathcal{W}^{\mathrm{cs}}$ is f–minimal, and hence cannot happen.

This finishes the proof of the lemma. □

5 Center dynamics in fixed leaves

This section deals with the dynamics of center leaves within center stable (and center unstable) leaves. It is one of the first places where we encounter significant difficulties compared with the dynamically coherent setting.

In [3, Proposition 4.4] we found a condition for the existence of center leaves that are fixed by a good lift, but the proof does not work without dynamical coherence [3, Remark 4.8].

Throughout this section we continue to assume that f is orientable (Definition 3.5).

Definition 5.1 Let $c \subset M$ be a center leaf that is fixed by f. We say that c is *coarsely contracting* if it is homeomorphic to the line, and it contains a nonempty compact interval I such that each compact interval $J \subset c$ whose interior contains I has the property that $f(J) \subset \mathring{J}$.

We say that c is called *coarsely expanding* if it is coarsely contracting for f^{-1}.

We also naturally extend the definition of coarse contraction/expansion to leaves that are periodic under f.

The following is the main result of this section.

Proposition 5.2 *Suppose that $\mathcal{W}^{cs}$ is f–minimal, and that there is a good lift $\widetilde{f}$ that fixes every center stable leaf but no center leaf in $\widetilde{M}$. Then every f–periodic center leaf in M is coarsely contracting.*

Note that a coarsely contracting periodic leaf must contain a periodic point.

If $\mathcal{W}^{cu}$ is f–minimal, and there is a good lift $\widetilde{f}$ that fixes every center unstable leaf in $\widetilde{M}$, then one concludes that each periodic center leaf is coarsely expanding.

We will see in Proposition 5.6 that one can always find f–periodic center leaves.

5.1 Fixed centers or coarse contraction

We begin with a preliminary result.

Lemma 5.3 *Suppose that $\widetilde{f}$ fixes every center stable leaf but no center leaf in $\widetilde{M}$. Then the same holds for every iterate $\widetilde{f}^n$ with $n \neq 0$.*

Proof Suppose that $\widetilde{f}^n$ fixes a center leaf c_0 for $n > 0$, and let L be a center stable leaf that contains c_0 (which is fixed by $\widetilde{f}$ by hypothesis). Since f is orientable, $\widetilde{f}$ preserves transverse orientations to the center and stable foliations on L.

Let A^{c} be the axis for the action of $\widetilde{f}$ on the center leaf space in L, ie the set of center leaves c such that $\widetilde{f}(c)$ separates c from $\widetilde{f}^2(c)$; see [3, Appendix E]. Since $\widetilde{f}^n(c_0)=c_0$, the leaf c_0 cannot be in A^{c}. If c_0 is not in ∂A^{c} then we can replace it with the unique center leaf that separates c_0 from A^{c}. Thus we can assume that $c_0 \in \partial A^{\mathrm{c}}$.

Recall (see [1, Proposition 2.15]) that the boundary of the axis of a homeomorphism on a 1–manifold splits into three disjoint sets: the "positive" boundary, "negative" and "middle" boundary. That is, ∂A^{c} contains three types of leaves, the center leaves c such that c and $\widetilde{f}(c)$ are nonseparated on their positive side, the leaves c such that c and $\widetilde{f}(c)$ are nonseparated on their negative side, and the leaves c that are nonseparated with a leaf c' in A^{c}.

If c_0 was in the "middle" boundary, then we would have that there exists $c' \in A^{\mathrm{c}}$ not separated with c_0. Thus c' and $\widetilde{f}^n(c')$ are separated, contradicting that $c_0=\widetilde{f}^n(c_0)$. So c_0 must be either in the positive or negative boundary. In particular, c_0 and $\widetilde{f}(c_0)$ are nonseparated.

Now, consider the closure of the set of stable leaves intersect c_0. There exists a unique stable leaf s_0 in the boundary of that set that separates c_0 from $\widetilde{f}(c_0)$. Therefore, s_0 must be fixed by $\widetilde{f}^n$, and hence contains a fixed point x of $\widetilde{f}^n$.

In particular, we found a periodic point of $\widetilde{f}$; thus, by the Brouwer translation theorem (see eg [2]) $\widetilde{f}$ must also admit a fixed point, say y. Since the center leaves through y form a closed interval (Lemma 3.19), there exists at least one closed center leaf through y, a contradiction. □

In order to obtain coarsely contracting center leaves we will use the following tool.

Proposition 5.4 *Suppose that $\widetilde{f}$ fixes every center stable leaf in $\widetilde{M}$, and let L be a center stable leaf that is also fixed by some $\gamma \in \pi_1(M) \setminus \{\mathrm{Id}\}$.*

Assume that there exists a properly embedded C^1–curve $\widehat{\eta} \subset L$ that is transverse to the stable foliation and fixed by both γ and $\widetilde{f}$.

- *If $\widetilde{f}$ does not act freely on $\mathcal{L}^{\mathrm{c}}_L$, then there is a center leaf in L fixed by both $\widetilde{f}$ and γ.*
- *If $\widetilde{f}$ acts freely on $\mathcal{L}^{\mathrm{c}}_L$, then every f–periodic center leaf in $\pi(L)$ is coarsely contracting.*

In the first case the center leaf projects to an f–invariant closed center leaf.

Note also that the hypotheses of Proposition 5.4 are implied by the conclusion of the graph transform lemma [3, Appendix H].

We will use the following result from [3], whose proof works equally well in the non-dynamically-coherent case.

Lemma 5.5 [3, Lemma 4.15] *Let c be a center leaf in a center stable leaf $L \subset \widetilde{M}$. Suppose that L is Gromov hyperbolic, and fixed by $\widetilde{f}$ and some nontrivial $\gamma \in \pi_1(M)$. Moreover, assume that there exist two stable leaves s_1, s_2 on L such that*

(1) *the center leaf c is in the region between s_1 and s_2,*

(2) *the leaves s_1 and s_2 are a bounded Hausdorff distance apart, and*

(3) *the leaves c, s_1 and s_2 are all fixed by $h = \gamma^n \circ \widetilde{f}^m$, $m \neq 0$.*

Then there is a compact segment $I \subset c$ such that h (if $m > 0$) or h^{-1} (if $m < 0$) acts as a contraction on $c \setminus \mathring{I}$.

Proof of Proposition 5.4 Since $\widetilde{f}$ fixes every leaf of $\mathcal{W}^{cs}$, Lemma 4.10 implies that it fixes no point in $\widetilde{M}$, and hence fixes no stable leaf.

Let S be the stable saturation of the curve $\hat{\eta}$. Let $\alpha = \pi(\hat{\eta})$. The curve α is closed, f–invariant, and tangent to the center bundle.

Case 1 We start by assuming that $\widetilde{f}$ fixes a center leaf c in L.

Suppose that c and $\hat{\eta}$ do not intersect a common stable leaf. Then c does not intersect the set S and there is a unique stable leaf s contained in the boundary of S such that s separates S from c. Since both S and c are $\widetilde{f}$–invariant, so is s. But then $\widetilde{f}$ must admit a fixed point in s, a contradiction.[3]

Therefore there is a stable leaf s intersecting c in y and $\hat{\eta}$ in x. Iterating forward by $\widetilde{f}$, we deduce that $d(\widetilde{f}^n(y), \widetilde{f}^n(x))$ converges to zero as y and x are in the same stable leaf. Since both c and $\hat{\eta}$ are $\widetilde{f}$–invariant, it implies that c and $\hat{\eta}$ are asymptotic; note that c and $\hat{\eta}$ may or may not intersect. Calling $\alpha = \pi(\hat{\eta})$ the projection to M, we deduce that $\pi(c)$ accumulates onto α. But as α is closed and $\pi(c)$ is a center leaf, we deduce that α is also a center leaf. Hence $\hat{\eta}$ is the required center leaf of the first option of the proposition.

[3]Note the distinction of c being fixed by $\widetilde{f}$ as opposed to $\pi(c)$ being periodic under f. It is the first property which creates a fixed point of $\widetilde{f}$ and a contradiction.

Case 2 Assume now that $\widetilde{f}$ acts freely on the center leaf space of L.

According to Lemma 5.3, $\widetilde{f}^n$ also acts freely on the center leaf space of L for any $n \neq 0$.

We need to prove now that every center leaf in $\pi(L)$ that is periodic must be coarsely contracting.

Let then c be a center leaf in L such that $\pi(c) = e$ is periodic under f, say of period m. Then, for some $\gamma_1 \in \pi_1(M) \setminus \{\mathrm{Id}\}$, we have $c = \gamma_1 \widetilde{f}^m(c)$. (One can show under our current assumptions that $\pi(L)$ projects to an annulus, so γ and γ_1 are both powers of a particular deck transformation, but we do not need that fact for the proof.) Let

$$h := \gamma_1 \circ \widetilde{f}^m.$$

We now want to show that either c intersects $\widehat{\eta}$, or there exists another center leaf, also fixed by h, that does.

Suppose first that c intersects S, ie there exists a stable leaf intersecting both c and $\widehat{\eta}$. Since the stable distance is contracted by h, which fixes both c and $\widehat{\eta}$, we obtain that either c and $\widehat{\eta}$ are asymptotic, or they intersect. If c and $\widehat{\eta}$ are asymptotic, then, as in Case 1, we deduce that $\widehat{\eta}$ must be a center leaf, contradicting the fact that $\widetilde{f}$ acts freely on the center leaf space. Thus we must have that c intersects $\widehat{\eta}$.

Suppose now that c does not intersect $\widehat{\eta}$, and thus does not intersect S. Then there is a unique stable leaf s in ∂S that separates $\widehat{\eta}$ from c. That leaf s must then be invariant by h, so admits a fixed point for h. Then at least one center leaf, say c_1, through that fixed point must be fixed by h. Since c_1 intersects S and is invariant by h, it must intersect $\widehat{\eta}$.

Thus in any case, we have a center leaf c_1 that intersects $\widehat{\eta}$, is invariant by h and, by the above argument, has both ends that escape compacts sets of L.

Let I be the projection of c_1 onto $\widehat{\eta}$ along stable leaves.

Suppose first that I is unbounded. Then, considering iterates by f^m, we deduce that $\pi(c_1)$ must be asymptotic to $\pi(\widehat{\eta})$, so $\widehat{\eta}$ must be a center leaf, which is not allowed, since $\widetilde{f}$ is assumed to act freely on center leaves.

So I is bounded in $\widehat{\eta}$. Let s_1 and s_2 be the stable leaves through the two endpoints of the interval I. Since I is fixed by h, so are s_1 and s_2. Moreover, the center leaf c_1, as well as c if it is different from c_1, is in between s_1 and s_2.

Now, $\widetilde{f}$ acts as a translation on $\widehat{\eta}$, so there exists $k \in \mathbb{Z}$ such that s_2 separates s_1 from $\widetilde{f}^k(s_1)$. By Lemma 4.17, s_1 and $\widetilde{f}^k(s_1)$ are a bounded Hausdorff distance apart. Thus s_1 and s_2 are a bounded Hausdorff distance apart. So c satisfies all the conditions for Lemma 5.5 to hold, thus it is coarsely expanding.

This finishes the proof of Proposition 5.4. □

We are now ready to prove the main result of this section.

Proof of Proposition 5.2 Let $e \subset M$ be an f–periodic center leaf, and let $c \subset \widetilde{M}$ be a lift of e. If $m > 0$ is the period of e, then c and $\widetilde{f}^m(c)$ both project to e, so there is an element $\gamma' \in \pi_1(M)$ with $\gamma'(\widetilde{f}^m(c)) = c$.

Choose a leaf $L \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$ that contains c. Then γ' is in the stabilizer of L, because $\widetilde{f}$ leaves invariant every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Since $\widetilde{f}^m$ acts freely on the center leaf space (cf Lemma 5.3), γ' is not the identity.

Since $\widetilde{f}$ does not have any fixed points, Proposition 4.15 implies that the stabilizer of L in $\widetilde{M}$ is infinite cyclic. Thus, there exists $\gamma \in \pi_1(M) \setminus \{\mathrm{id}\}$ such that $\gamma^n \circ \widetilde{f}^m(c) = c$ for some $n \in \mathbb{Z}$ with $n \neq 0$, and such that γ generates the stabilizer of L. Let

$$h := \gamma^n \circ \widetilde{f}^m.$$

Notice that h is still a partially hyperbolic diffeomorphism and has bounded derivatives.

Since $\widetilde{f}$ acts freely on $\mathcal{L}^{\mathrm{c}}_L$, it must also act freely on $\mathcal{L}^{\mathrm{s}}_L$. Let A^{s} be the axis for the action of $\widetilde{f}$ on the stable leaf space $\mathcal{L}^{\mathrm{s}}_L$; see [3, Appendix E]. No stable leaf in M can be closed, so γ must also act freely on $\mathcal{L}^{\mathrm{s}}_L$. Since γ and $\widetilde{f}$ commute, A^{s} is also the axis for the action of γ on $\mathcal{L}^{\mathrm{s}}_L$. The axis A^{s} can be a line or a countable union of intervals.

Suppose first that A^{s} is a line. Let s be a stable leaf in A^{s} and p in s. Then p and γp can be connected by a transversal to the stable foliation, chosen so that the projection to $\pi(L)$ is a smooth simple closed curve. Let η be the union of the γ iterates of this segment. Applying the graph transform lemma [3, Lemma H.1] to η we obtain a curve $\widehat{\eta}$ which satisfies the properties in the hypothesis of Proposition 5.4, as desired.

Now suppose that A^{s} is a countable union of intervals

$$A^{\mathrm{s}} = \bigcup_{i \in \mathbb{Z}} [s_i^-, s_i^+] = \bigcup_{i \in \mathbb{Z}} T_i.$$

Our first claim is that there exists $s \in A^{\mathrm{s}}$, fixed by h, such that the center leaf c is between $\gamma^{-1}s$ and γs.

Suppose that c intersects some stable leaf s' in A^{s}. Then s' is in a unique T_i for some i (the center leaf c cannot intersect two different intervals otherwise c would intersect two nonseparated leaves, which is impossible). Then, since h fixes c, it also fixes the axis A^{s} and preserves the transverse orientation. It follows that $h(T_j) = T_j$ for all j. In this case we set $s = s_i^+$. The leaf s is fixed by h and there exists $k \neq 0$ such that $\gamma^{\pm 1} T_i = T_{i \pm k}$. Thus T_i is in between $\gamma^{-1}s$ and γs and hence, so is c. Recall here that h preserves orientation.

Now, suppose instead that c does not intersect A^{s}. Hence, there is a unique i such that $s_{i-1}^+ \cup s_i^-$ separates c from all other stable leaves in A^{s}. We again set $s := s_i^+$. As before, since h fixes both c and A^{s}, and preserves the transverse orientation, it must fix s also. The same argument as above also shows that c is between $\gamma^{-1}s$ and γs.

In either case we have found a stable leaf s (chosen as a positive endpoint of one of the closed intervals T_i) that is fixed by h, such that c lies between $\gamma^{-1}s$ and γs. Notice that both γs and $\gamma^{-1}s$ are fixed by h.

The leaf $\gamma^{-1}s$ is between γs and $\widetilde{f}^{2m}(\gamma s) = \gamma^{-2n+1}s$ (assuming $n \geq 1$, otherwise between γs and $f^{-2m}(\gamma s)$). Hence the Hausdorff distance between $\gamma^{-1}s$ and γs is bounded above by a uniform constant $C > 0$, depending only on f and m.

Thus the center leaf c, fixed by h, lies between the stable leaves γs and $\gamma^{-1}s$, also fixed by h, which are a bounded Hausdorff distance apart. Moreover, the leaves of $\mathcal{W}^{\mathrm{cs}}$ are Gromov hyperbolic by Lemma 4.18. These are all the conditions needed to apply Lemma 5.5, so c is coarsely contracting for h. $\square$

5.2 Existence of periodic center leaves

In order to apply Propositions 5.2 and 5.4 we will need some way to find periodic center leaves.

Proposition 5.6 *Let $f : M \to M$ be a partially hyperbolic diffeomorphism homotopic to the identity.*

Suppose that some good lift $\widetilde{f}$ fixes every center stable leaf in $\widetilde{M}$. If L is a center stable leaf fixed by some $\gamma \in \pi_1(M) \setminus \{\mathrm{Id}\}$, then there is an f–periodic center leaf in $\pi(L)$.

Proof First notice that if one can prove the above result for a finite cover of M and a finite power of f, then the same result directly follows for the original map and manifold. Thus, we may assume that M is orientable, f is orientation-preserving, and the branching foliations are both transversely orientable.

Given these assumptions, L projects to an annulus in M (Proposition 4.15). Let γ be a generator of the stabilizer of L.

If $\widetilde{f}$ fixes a center leaf in L, then it would project to a center leaf fixed by f, proving the claim. So we assume that $\widetilde{f}$ acts freely on the center leaf space in L. This implies that $\widetilde{f}$ also acts freely on the stable leaf space in L, and we can thus consider the stable axis $A \subset \mathcal{L}^s_L$ of $\widetilde{f}$. Since γ also acts freely on the stable leaves, and commutes with $\widetilde{f}$, it has the same set A as its axis. This axis is either a line or a countable union of intervals.

If the axis is a countable union of intervals, there must be integers n, m such that $h := \gamma^n \widetilde{f}^m$ fixes one of the intervals, and hence a stable leaf s. One cannot have $m = 0$, since this would mean that γ^n would fix a stable leaf, which is impossible. So $m \neq 0$, and s projects to a periodic stable leaf $\pi(s)$ in M. This must contain a periodic point, and at least one center leaf through that point is periodic, as desired.

If the axis is a line, then one can use the graph transform lemma [3, Appendix H] to see that there is a properly embedded curve in L which is invariant under $\widetilde{f}$ and γ. Then [3, Lemma H.3] provides a periodic center leaf, as desired. □

5.3 Additional result

The intermediate results in this section also provide a proof of the following result, which will be needed later in this article.

Proposition 5.7 *Suppose that $\widetilde{f}$ fixes every center stable leaf in $\widetilde{M}$, and let L be a center stable leaf that is also fixed by some $\gamma \in \pi_1(M) \setminus \{\mathrm{Id}\}$. Assume moreover that there is no center leaf in L fixed by $\widetilde{f}$. Then there is a center leaf c in L fixed by $h = \gamma^n \circ \widetilde{f}^m$ for some n, m with $m \neq 0$, and two stable leaves s_1, s_2 on L such that*

(1) *the center leaf c separates s_1 from s_2 in L,*

(2) *the leaves s_1 and s_2 are a bounded Hausdorff distance apart, and*

(3) *the leaves c, s_1 and s_2 are all fixed by $h = \gamma^n \circ \widetilde{f}^m$, where $m \neq 0$.*

Proof The conditions imply that $\pi(L)$ is an annulus. Proposition 5.6 implies that there is a periodic center in $\pi(L)$.

To prove Proposition 5.7 we revisit the proof of Proposition 5.2. Since there is no center fixed by $\widetilde{f}$ in L, then as in the proof of Proposition 5.2 the map $\widetilde{f}$ acts freely on the stable leaf space. As in that proposition we separate into whether the stable axis is a line or a $\mathbb{Z}$–union of intervals.

In the first case, as in Proposition 5.2 we produce a curve $\widehat{\eta}$ in L which is invariant under $\widehat{f}$ and γ. We will use Proposition 5.4, and the existence of such a curve $\widehat{\eta}$ is necessary for that. The analysis of Proposition 5.4 has cases depending on the action of $\widetilde{f}$ on the center leaf space — as opposed to the action on the stable leaf space A^{s}. However, in this proposition we are assuming that the action on the center leaf space in L is free, so this is Case 2 of Proposition 5.4, where the proof showed the existence of a center leaf c and stable leaves s_1, s_2 satisfying the conditions stated in this proposition, except perhaps that c separates s_1 from s_2.

We now show that such a center leaf exists with this additional property. Suppose that this does not happen for c. This can only occur if both ends of $\pi(c)$ are in the same end of the annulus $\pi(L)$, or in other words, if $\pi(c)$ separates $\pi(L)$. Since the action of $\widetilde{f}$ on the center leaf space in L is free it has an axis denoted by A^{c}. The leaf c is not in this axis. If the axis A^{c} is a real line then there is a unique center leaf c' in the axis A^{c} which is either nonseparated from c or is nonseparated from a leaf which separates c from the axis. In either case it also follows that h fixes c'. We can then redo the analysis with c' instead of c. It will produce stable leaves s_1, s_2 fixed by h, with c' between them, and now c' separates s_1 from s_2. If the center axis A^{c} is a countable union of intervals, there is a unique consecutive pair of intervals such that c is "between" them. Then the boundary leaves of these intervals are fixed by h. Choose c' to be one of these boundary leaves, and redo the proof with c' instead of c to obtain the conclusion of the proposition.

The other case of this proposition is when the stable axis is a $\mathbb{Z}$–union of intervals. Here we use the notation of the proof of Proposition 5.2, where $A^{\mathrm{s}} = \bigcup_{i \in \mathbb{Z}} [s_i^-, s_i^+] = \bigcup_{i \in \mathbb{Z}} T_i$. Consider s_0^+, which is nonseparated in the stable leaf space from s_1^-. There are n, m, with $m \neq 0$, such that $h = \gamma^n \circ \widetilde{f}^m$ fixes all T_i and their boundary leaves. Since s_0^+ and s_1^- are nonseparated, consider a nearby stable leaf s which intersects transversals to both of them. Choose a center c_0 intersecting s and s_0^+, and choose a center c_1 intersecting s and s_1^-. Starting from c_0 and considering the centers intersecting s between $c_0 \cap s$ and $c_1 \cap s$, there is a first center leaf, denoted by c, which does not intersect s_0^+. This center is fixed by h. Let $s_1 = s_0^+$ and $s_2 = s_1^-$. They are both fixed by h. In addition, c separates s_1 from s_2. Finally, s_1 and s_2 are a finite Hausdorff distance from each other in L.

This completes the proof of the proposition. $\square$

6 Minimality for Seifert and hyperbolic manifolds

In this section we will show that when M is hyperbolic or Seifert, the existence of a single fixed center stable leaf implies that every center stable leaf is fixed. This is considerably easier in the dynamically coherent case [3, Proposition 3.15].

We continue to assume that f is orientable.

Proposition 6.1 *Suppose that M is hyperbolic or Seifert fibered, and a good lift $\widetilde{f}$ fixes some leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Then $\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, $\mathcal{W}^{\mathrm{cs}}$ is f–minimal, and every leaf of $\mathcal{W}^{\mathrm{cs}}_{\epsilon}$ and $\mathcal{W}^{\mathrm{cs}}$ is either a plane or an annulus. The same statement holds for $\mathcal{W}^{\mathrm{cu}}$.*

The main issue in extending the proof of [3] to the non-dynamically-coherent context is that here we cannot ensure the nonexistence of fixed points of $\widetilde{f}$, since Lemma 4.10 does not exclude fixed points when the branching foliation is not f–minimal. So we will need to exclude the existence of fixed points for good lifts. We cannot exclude their existence in general, but we are able to show that they cannot exist in minimal sublaminations of $\mathcal{W}^{\mathrm{cs}}$ or $\mathcal{W}^{\mathrm{cu}}$.

6.1 No fixed points for good lifts

Note that the definition of f–minimality for the whole foliation can be applied to subsets: a $\mathcal{W}^{\mathrm{cs}}$–saturated subset of M is *f–minimal* if it is closed, nonempty, and f–invariant, and no proper saturated subset satisfies these conditions.

The goal of this subsection is to prove the following proposition, which does not assume that M is hyperbolic or Seifert.

Proposition 6.2 *Let $\widetilde{f}$ be a good lift of f to $\widetilde{M}$. Suppose that Λ is a nonempty f–minimal set of $\mathcal{W}^{\mathrm{cs}}$ such that every leaf L of $\widetilde{\Lambda} = \pi^{-1}(\Lambda)$ is fixed by $\widetilde{f}$. Then $\widetilde{f}$ has no fixed points in $\widetilde{\Lambda}$.*

We will prove this proposition by contradiction. So from now on, we assume that there is a fixed point p of $\widetilde{f}$ in a leaf L contained in $\widetilde{\Lambda}$. This point projects to a fixed point $\pi(p)$ in M. Note that any leaf L' of $\widetilde{\Lambda}$ that intersects the unstable leaf $u(p)$ through p must have $L' \cap u(p) = p = L \cap u(p)$. This is because L and L' are both fixed, and unstable leaves are expanded.

6.1.1 Many fixed points The following property uses crucially the fact that Λ is an f–minimal sublamination.

Lemma 6.3 *There exists $b > 0$ such that any point in a leaf of $\widetilde{\Lambda}$ is at distance at most b (for the path metric on the leaf) from a fixed point of $\widetilde{f}$.*

Proof Otherwise, one can find a sequence of discs D_i in leaves of $\widetilde{\Lambda}$ that contain no fixed points and whose radius goes to ∞. Up to deck transformations and subsequences, these disks converge to a full leaf L_1 of $\widetilde{\mathcal{W}}^{cs}$ that is contained in $\widetilde{\Lambda}$. Here, the convergence is with respect to the topology of the center stable leaf space, which also implies convergence as a set of $\widetilde{M}$. The leaf L_1 does not contain any fixed point of $\widetilde{f}$. Indeed, the unstable leaf through a fixed point q in L_1 would eventually intersect one of the discs D_i. Since $\widetilde{f}$ fixes the leaves of $\widetilde{\Lambda}$, this would imply that the leaf through D_i merges with L_1 and that D_i contains the fixed point q, a contradiction.

Since L_1 contains no fixed points, it does not contain the $\widetilde{f}$–fixed point p, and $A = \pi(L_1)$ does not contain the f–fixed point $\pi(p)$. But the closure of $A = \pi(L_1)$ in M is Λ by minimality, so A must accumulate on $\pi(p)$. But this means that A intersects $u(\pi(p))$, which implies that A contains $\pi(p)$ as explained above. This is a contradiction. □

6.1.2 A topological lemma Let L be a metrically complete, noncompact, simply connected, Riemannian surface.

For a compact subset $X \subset L$ we denote by $\mathrm{Fill}(X)$ the complement of the unique unbounded connected component of $L \setminus X$. Note that $\mathrm{Fill}(X)$ is always compact, as a neighborhood of ∞ in the compactification of L is disjoint from X. Notice further that, by definition, $\mathrm{Fill}(X)$ is a compact connected set which does not separate the plane.

We will use the following simple properties of $\mathrm{Fill}(X)$:

- If $X \subset Y$ are compact sets, then $\mathrm{Fill}(X) \subset \mathrm{Fill}(Y)$.
- If $g\colon L \to L$ is a homeomorphism and $X \subset L$, then $g(\mathrm{Fill}(X)) = \mathrm{Fill}(g(X))$.

The following lemma will be used in the next section; see Figure 5.

Lemma 6.4 *Let L be as above. Then for every $b > 0$ and $\delta > 0$ there exists $R > 0$ and $n_0 > 0$ with the following property. Let A and B be topological disks, and let $\ell_1, \dots, \ell_n$, with $n \geq n_0$, be disjoint curves that join A and B. Suppose that*

(i) *$d(A, B) > 2R$, and*

(ii) *the δ–neighborhoods of the curves ℓ_i are pairwise disjoint.*

Then the region $\mathrm{Fill}(A \cup B \cup \ell_1 \cup \cdots \cup \ell_n) \setminus (A \cup B)$ contains a disk D of radius $> 4b$. Moreover, D can be chosen so that $d(D, A)$ and $d(D, B)$ are larger than $d(A, B)/10$.

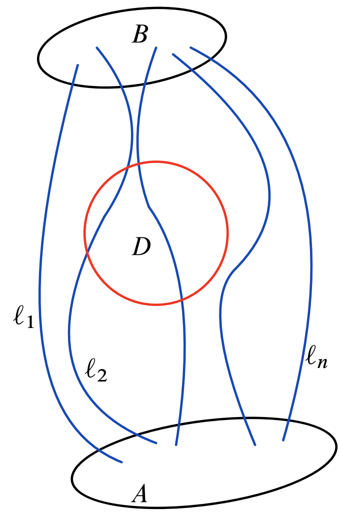

Figure 5: A depiction of Lemma 6.4.

Proof Using the Jordan curve theorem we can reorder the curves so that

- $\operatorname{Fill}(A \cup B \cup \ell_1 \cup \cdots \cup \ell_n) = \operatorname{Fill}(A \cup B \cup \ell_1 \cup \ell_n)$, and
- for $1 < i < n$ we have that $\ell_i \subset \operatorname{Fill}(A \cup B \cup \ell_{i-1} \cup \ell_{i+1})$.

Take $R > 100b$ and $n_0 > 100b/\delta$. Without loss of generality we can assume that n is even. This way, we can choose a point $x \in \ell_{n/2}$ such that $d(x, A) > d(A, B)/4$ and $d(x, B) > d(A, B)/4$. We claim that $B(x, 4b) \subset \operatorname{Fill}(A \cup B \cup \ell_1 \cup \ell_n)$. By our choice of x it will follow that $B(x, 4b)$ is at distance larger than $d(A, B)/10$ from A and B.

To see this, consider a geodesic ray r starting from x, and let y be the first point of intersection of r with $\partial\operatorname{Fill}(A \cup B \cup \ell_1 \cup \cdots \cup \ell_n) \setminus (A \cup B)$. By our ordering, there are two possibilities: either

- y belongs to $\partial A \cup \partial B$, or
- y belongs to $\ell_1 \cup \ell_n$.

By our assumptions, if $y \in \partial A \cup \partial B$ then the distance $d(x, y) > R/4 > 4b$. On the other hand, if $y \in \ell_1$ then by our choice of reordering we deduce that r must intersect ℓ_i for all $1 \le i \le n/2$. Since the points of intersection are at distance pairwise larger than δ, we deduce that $d(x, y) > 4b$. Similarly, if $y \in \ell_n$ we also get $d(x, y) > 4b$. This completes the proof. $\square$

6.1.3 Proof of Proposition 6.2 We will repeatedly use the fact that $\widetilde{f}^{-1}$ expands stable length. To simplify notation we set $g := \widetilde{f}^{-1}$. The rest of this subsection is devoted to the proof of Proposition 6.2.

According[4] to Lemma 4.17, there is a constant $K > 0$ such that, for any $z \in L$, we have

$$d_L(z, g(z)) \leq K,$$

where d_L denotes the path metric on L. From now on within this subsection we will always work in L, so we will drop the subscript and write $d := d_L$.

To finish the proof, our aim will be to show that the fact that $\widetilde{f}$ moves points a bounded distance in L, together with the exponential contraction of length along the stable leaf $s(p)$ under iteration by $\widetilde{f}$, will force an arbitrarily large amount of "bunching" of $s(p)$, which is impossible for leaves of planar foliations.

Indeed, since $s(p)$ is a leaf of a foliation of the plane, there exist some constants $\delta, \eta > 0$ such that if $I, J \subset s(p)$ are closed segments which are at distance larger than η in the $s(p)$ metric, then their δ–neighborhoods are disjoint in L. Now, this implies in particular that the volume of the δ–neighborhood of a segment of $s(p)$ must grow to infinity with its length (and therefore, the diameter grows to infinity with the length).

Without loss of generality, we can assume that $\delta, \eta < 1$ and $K > 1$.

To prove Proposition 6.2 we will fix $b > 0$ as given by Lemma 6.3, and $\delta > 0$ by the considerations above. Lemma 6.4 then gives us values of $R > 0$ and $n_0 > 0$ associated to b and δ so that its statement holds. We will fix

$$n > \max\left\{\frac{10R}{K}, \frac{10b}{\delta}, n_0\right\}.$$

We introduce the following notation: given any $a, b \in s(p)$, we write $[a,b]^{\mathrm{s}}$ to indicate the closed segment along the stable leaf $s(p)$ between a and b, oriented from a to b.

We will then pick points in $y, z \in s(p)$ with the properties

- $d(y,z) > 200Kn$,
- $g([y,z]^{\mathrm{s}}) \cap [y,z]^{\mathrm{s}} = \varnothing$ (equivalently, $z \in [y, g(y)]^{\mathrm{s}}$).

The existence of points like this follows from the fact that if y_0 is any point in $s(p)$, the length of $g^k([y_0, g(y_0)]^{\mathrm{s}})$ grows to infinity, and thus its diameter grows too. See Figure 6.

[4]It is not hard to see that the proof applies to the fixed sublamination.

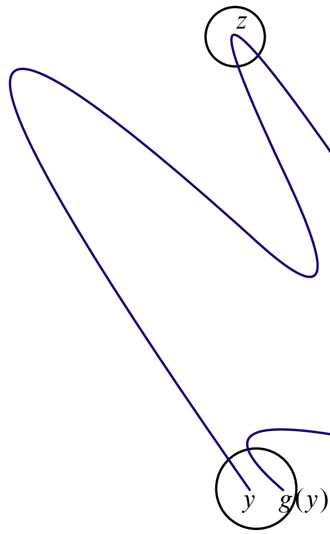

Figure 6: Choosing points y and z in $s(p)$.

We will pick $A_i = B(y, Ki)$ and $B_i = B(z, Ki)$, the neighborhoods of radius Ki of the points y and z. Given our choices, notice that $g(y) \in A_1$, $g(z) \in B_1$, and, for any i, we have $g(A_i) \subset A_{i+1}$ as well as $g(B_i) \subset B_{i+1}$.

The following holds:

Lemma 6.5 *Every arc $J \subset [y, g^n(z)]^{\mathrm{s}}$ which is disjoint from $A_n \cup B_n$ is completely contained in a fundamental domain of $s(p)$ for the action of g. More precisely, there exists ℓ such that $J \subset [g^\ell(y), g^\ell(z)]^{\mathrm{s}}$ or $J \subset [g^\ell(z), g^{\ell+1}(y)]^{\mathrm{s}}$.*

Proof This is because $[y, z]^{\mathrm{s}}$ intersects A_1 and B_1, so every fundamental domain as above intersects both A_n and B_n. □

We can apply Lemma 6.4 to get:

Lemma 6.6 $\mathrm{Fill}(A_n \cup B_n \cup [y, g^n(z)]^{\mathrm{s}}) \setminus (A_{10n} \cup B_{10n})$ *contains a disk of radius* $4b$.

Proof Note that $[g^\ell(y), g^{\ell+1}(y)]^{\mathrm{s}}$ contains at least two segments joining A_{10n} to B_{10n} if $0 \le \ell < n$; see Figure 6. Thus there are at least $2n$ such curves, which, since they are segments separated in $s(p)$, must have pairwise disjoint δ–neighborhoods. Thus, by our choice of constants b, δ, K and n above, we can apply Lemma 6.4 to find a disk of radius $\geq 4b$ inside $\mathrm{Fill}(A_n \cup B_n \cup [y, g^n(z)]) \setminus (A_n \cup B_n)$ which is at distance larger than $d(A_n, B_n)/10$ from A_n and B_n. Thus, the disk is contained in $\mathrm{Fill}(A_n \cup B_n \cup [y, g^n(z)]^{\mathrm{s}}) \setminus (A_{10n} \cup B_{10n})$, as required. □

Using Lemma 6.3 we can find a fixed point $q \in \mathrm{Fix}(g)$ such that

$$B(q, 2b) \subset \mathrm{Fill}(A_n \cup B_n \cup [y, g^n(z)]^{\mathrm{s}}) \setminus (A_{10n} \cup B_{10n}).$$

We can show the following:

Lemma 6.7 *There exists an arc $J \subset [y, g^n(z)]^{\mathrm{s}}$ such that either*

(1) *J intersects A_n only at its endpoints and $q \in \mathrm{Fill}(A_n \cup J)$, or*

(2) *J intersects B_n only at its endpoints and $q \in \mathrm{Fill}(B_n \cup J)$.*

Moreover, J is contained in a fundamental domain: for some $0 \leq \ell \leq n$ we either have $J \subset [g^\ell(y), g^{\ell+1}(y)]^{\mathrm{s}}$ or $J \subset [g^\ell(z), g^{\ell+1}(z)]^{\mathrm{s}}$.

Proof This follows from the fact that $\mathrm{Fill}(A_n \cup B_n \cup [y, g^n(z)]^{\mathrm{s}})$ is contained in a union of sets of this form.

To see this, note that

$$\mathrm{Fill}(A_n \cup B_n \cup [y, g^n(z)]^{\mathrm{s}}) = \mathrm{Fill}(A_n \cup [y, g^n(z)]) \cup \mathrm{Fill}(B_n \cup [y, g^n(z)]^{\mathrm{s}})$$

because A_n and B_n are disjoint topological disks and $[y, g^n(z)]^{\mathrm{s}}$ is a topological interval. Indeed, by Jordan's theorem $\widehat{A} = \mathrm{Fill}(A_n \cup [y, g^n(z)]^{\mathrm{s}})$ is a topological disk with an arc attached (ie the segment $[w, g^n(z)]^{\mathrm{s}}$ where w is the last point of intersection of $[y, g^n(z]]^{\mathrm{s}})$, and similarly $\widehat{B} = \mathrm{Fill}(B_n \cup [y, g^n(z)]^{\mathrm{s}})$ is a topological disk with an arc attached. One has that $\mathrm{Fill}(A_n \cup B_n \cup [y, g^n(z)]^{\mathrm{s}}) = \mathrm{Fill}(\widehat{A} \cup \widehat{B})$. Since the intersection of these sets is connected (because their intersection retracts to $[y, g^n(z)]^{\mathrm{s}}$) we deduce[5] that $\mathrm{Fill}(\widehat{A} \cup \widehat{B}) = \widehat{A} \cup \widehat{B}$.

The fact that J is contained in a fundamental domain is a direct consequence of the fact that it intersects A_n (or B_n) only in its boundaries, and thus Lemma 6.5 can be applied. □

[5] Here we are using the fact from plane topology that generalizes the Jordan curve theorem stating that if X and Y are compact connected sets, then their union separates the plane if and only if their intersection is not connected.

Both cases are analogous, so we will assume from now on that the first option happens, namely, $q \in \mathrm{Fill}(A_n \cup J)$ for a curve $J \subset [y, g^n(z)]^{\mathrm{s}}$ which intersects A_n only at its endpoints and such that J is contained in a fundamental domain of $s(p)$.

To reach a contradiction, the idea will be to find fixed points q_1, q_2 which are sufficiently close, and such that one belongs to $\mathrm{Fill}(A_n \cup J)$ and the other does not. If we choose them appropriately, we will be able to see that $g^i(J)$ will intersect a geodesic joining q_1 and q_2 for several values of i (before the set $g^i(A_n)$ becomes too big). This will produce some accumulation of the arcs $g^i(J)$ (which are segments of $s(p)$ far along $s(p)$); this is not possible, and gives the desired contradiction.

Lemma 6.8 *There are fixed points $q_1, q_2 \in \mathrm{Fix}(g)$ such that $d(q_1, q_2) < 3b$ and we have that $q_1 \in \mathrm{Fill}(A_n \cup J) \setminus A_{10n}$ while $q_2 \notin \mathrm{Fill}(A_n \cup J)$.*

Proof We will use Lemma 6.3. By the choice of the point q we can consider an unbounded geodesic ray r starting at q which is at distance larger than $2b$ from A_{10n}. One can cover r by balls of radius b; in each such ball there is a fixed point, and eventually, the fixed point is not in $\mathrm{Fill}(A_n \cup J)$, which is a bounded set. So there is a pair of such points for which one belongs to $\mathrm{Fill}(A_n \cup J)$ and the other does not. Their distance is less than $3b$. □

We are now ready to prove Proposition 6.2 by finding a contradiction, which will be produced using the following:

Lemma 6.9 *For every $0 \le i \le n$, we have that $g^i(J) \cap [q_1, q_2]_L \neq \varnothing$, where $[q_1, q_2]_L$ denotes a geodesic segment joining q_1 and q_2.*

Proof Note first that since $d(q_1, q_2) < 3b$ and $q_1 \notin A_{10n}$, we know that the geodesic segment $[q_1, q_2]_L$ is disjoint from A_{5n} (recall that $\delta < 1$ and that $n > 10b/\delta$).

Since $q_1 \in \mathrm{Fill}(A_n \cup J)$ is fixed, we get that $q_1 = g^i(q_1) \in g^i(\mathrm{Fill}(A_n \cup J)) = \mathrm{Fill}(g^i(A_n) \cup g^i(J))$. Similarly, we get that since $q_2 \notin \mathrm{Fill}(A_n \cup J)$, we have that $q_2 \notin \mathrm{Fill}(g^i(A_n) \cup g^i(J))$.

This implies that $\partial \mathrm{Fill}(g^i(A_n) \cup g^i(J))$ must intersect $[q_1, q_2]_L$. Since $g^i(A_n) \subset A_{n+i}$, which is disjoint from $[q_1, q_2]_L$, we deduce that $g^i(J)$ must intersect $[q_1, q_2]_L$, as we wanted to show. □

The contradiction is now the fact that $g^i(J)$ are curves whose δ–neighborhoods are disjoint, and all intersect $[q_1, q_2]_L$, which is a geodesic segment of length $< 3b$. This produces n different points at pairwise distance $\geq \delta$ in $[q_1, q_2]_L$, which is a contradiction since $n > 10b/\delta$.

6.2 Proof of Proposition 6.1

We are now ready to prove Proposition 6.1.

This proof follows the same structure as the one of [3, Proposition 3.15] and we will continuously refer to it. Recall the standing assumption that f is orientable.

Consider Λ an f–minimal nonempty subset. We need to show that $\Lambda = M$. We assume for the sake of contradiction that $\Lambda \neq M$.

Since $\mathcal{W}^{\mathrm{cs}}$ has no closed leaves and Λ is f–minimal, there cannot be any isolated leaves in Λ (for the topology of the stable leaf space).

Now, Proposition 6.2 allows us to assert that $\widetilde{f}$ has no fixed points in leaves of $\widetilde{\Lambda}$. Then, Corollary 6.12 implies that each leaf of Λ is either a plane or an annulus.

Fix an ϵ small enough and let Λ' be the pullback of Λ to the approximating foliation $\mathcal{W}^{\mathrm{cs}}_\epsilon$. That is, $\Lambda' = (h^{\mathrm{cs}}_\epsilon)^{-1}(\Lambda)$. Let V be a connected component of $\widetilde{M} \setminus \widetilde{\Lambda}'$.

Claim 6.10 *The projection $\pi(V)$ to M has finitely many boundary leaves.*

This is a standard fact in the theory of foliations [14, Lemma 5.2.5].

Claim 6.11 *Each leaf $L \subset \partial V$ projects to an annulus $\pi(L)$ in M.*

Proof Suppose that $\pi(L)$ is a plane. Recall (see [14, Lemma 5.2.14]) that $\pi(V)$ has an octopus decomposition and a compact core. So for any $\delta > 0$, the subset of points in $\pi(L)$ that are at distance greater than δ from another boundary component of $\pi(V)$ is precompact. Since $\pi(L)$ is supposed to be a plane, that subset must be contained in a closed disk D. Then $\pi(L) \setminus D$ is an annulus that is δ–close to another boundary component, $\pi(L')$, of $\pi(V)$. Moreover, the subset of $\pi(L')$ that is δ–close to $\pi(L) \setminus D$ then also has to be an annulus. If $\pi_1(L')$ were not a plane it would be an annulus and its nontrivial curve corresponds to a curve homotopic to the boundary of the closed disk D, which is homotopically trivial in M. Since the leaves of $\mathcal{W}^{\mathrm{cs}}_\epsilon$ are π_1–injective, this implies that $\pi(L')$ is also a plane.

Since M is irreducible this implies that $\pi(V)$ is homeomorphic to an open disk times an interval. So $\pi(V)$ has only two boundary components, both of which are planes. In particular, the isotropy group of V is trivial and $\pi(V)$ is homeomorphic to V.

We will now switch to the branching foliation to finish the proof. Let $A = h^{cs}_\epsilon(\pi(L))$ and $B = h^{cs}_\epsilon(\pi(L'))$. Since we chose ϵ small enough, up to taking δ small enough also, the unstable segments through $A \setminus h^{cs}_\epsilon(D)$ intersect B, and their length is uniformly bounded. Moreover, no unstable ray of A can stay in $h^{cs}_\epsilon(\pi(V))$. This is because $\pi(V)$ is homeomorphic to an open disk times an interval. So, since D is compact, the length of every unstable segment between A and B is bounded by a uniform constant. Notice that, since $\mathcal{W}^{cs}$ is a branching foliation, we may have $A \cap B \neq \varnothing$, ie some of these unstable segments may be points.

Since L and L' are in ∂V, which is a connected component of $\widetilde{M} \setminus \widetilde{\Lambda}'$, we have that $A, B \in \partial(M \setminus \Lambda)$. So in particular, A and B are fixed by f. Hence, the set of unstable segments between A and B is also invariant by f. Since the lengths of unstable segments between A and B are bounded above and f expands the unstable length, all the unstable segments must have zero length, ie $A = B$. This implies that V is empty, which contradicts the assumption that $\Lambda \neq M$. □

Thus we showed that every component of $\pi(\partial V)$ is an annulus. We can then apply without change the (topological) arguments of the proof of [3, Proposition 3.15] to obtain a torus T, composed of annuli along leaves of $\mathcal{W}^{cs}_\epsilon$ together with annuli transverse to $\mathcal{W}^{cs}_\epsilon$, which bounds a solid torus U' in $\pi(V)$.

Now consider $U = h^{cs}_\epsilon(U')$. Because of the collapsing of leaves, U may not be a solid torus. If U is empty for any such component U', this would directly contradict the assumption $\Lambda \neq M$. So for some such complementary component U', the set U is not empty and it is contained in a solid torus (the ϵ–tubular neighborhood of U' in M). We can then use the same "volume vs length" argument on U, exactly as in the end of the proof of [3, Proposition 3.15], to get a final contradiction. This ends the proof of Proposition 6.1.

6.3 Some consequences

An important consequence of Proposition 6.2 is the following:

Corollary 6.12 *Suppose that f is a partially hyperbolic diffeomorphism in M that is homotopic to the identity. Let $\widetilde{f}$ be a good lift of f to $\widetilde{M}$. Suppose that Λ is a* *nonempty (saturated) f–minimal subset of $\mathcal{W}^{cs}$ such that every leaf of the lift $\widetilde{\Lambda}$ to $\widetilde{M}$ is fixed by $\widetilde{f}$. Then every leaf in the f–minimal set Λ of $\mathcal{W}^{cs}$ is either a plane or an annulus.*

Proof Let A be a leaf of Λ and L a lift in $\widetilde{M}$. By Proposition 6.2, L does not admit any fixed points of $\widetilde{f}$. Hence, $\widetilde{f}$ acts freely on the space of stable leaves in L.

Now, recall that $\pi_1(A)$ can be defined as the elements $\gamma \in \pi_1(M)$ that fix L; see Section 4.3. So if $\gamma \in \pi_1(A)$, it must also act freely on the space of stable leaves in L. As $\widetilde{f}$ commutes with every deck transformation, Corollary E.4 of [3] — which still applies in the context of branching foliation, as does all of [3, Appendix E] — implies that $\pi(A)$ is abelian, ie A is either a plane or an annulus (again with the understanding that A might actually only be an immersion of one of these manifolds in M, and recalling that all bundles were assumed to be orientable in this section, so in particular the leaves cannot be Möbius bands). $\square$

As a consequence, we also get the following result, which completes the proof of Theorem 1.3 as announced.

Corollary 6.13 *Suppose that f is a partially hyperbolic diffeomorphism homotopic to the identity. Suppose that f is either volume-preserving or transitive, or that M is either hyperbolic or Seifert. Let $\widetilde{f}$ be a good lift of f. Then $\widetilde{f}$ has no periodic points. In particular, f has no contractible periodic points.*

Proof Up to finite covers and iterates, we may assume that f preserves the branching foliations $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$.

If $\widetilde{f}$ acts as a translation on either $\mathcal{W}^{cs}$ or $\mathcal{W}^{cu}$, then it does not have periodic points.

Otherwise, since we showed that under our assumptions the branching foliations are f–minimal, the result then follows from Theorem 4.12. $\square$

7 Double invariance implies dynamical coherence

In this section we show that if the center stable and center unstable branching foliations are minimal and leafwise fixed by a good lift $\widetilde{f}\colon \widetilde{M} \to \widetilde{M}$, then, f has to be dynamically coherent (ie the branching foliations do not branch). Therefore, we will be able to apply the results from the dynamically coherent setting.

The universal cover $\widetilde{M}$ of M is homeomorphic to $\mathbb{R}^3$ (since it admits a partially hyperbolic diffeomorphism; see [3, Appendix B]). We do not assume anything further on M in this section.

Recall also that a center leaf is a connected component of the intersection of a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and one of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$; cf Definition 3.7.

This section (and the proof of dynamical coherence) is split into three parts. First, in Section 7.1, we show that for an appropriate lift of M and power of f, double invariance of the foliations implies that the center leaves are fixed. The lift and power we need to consider here is in order to have everything orientable and coorientable. Then, in Section 7.2, we show that if a good lift fixes every center leaf, then it must be dynamically coherent. Finally, in Section 7.3, we show that if a lift and power of a partially hyperbolic diffeomorphism is dynamically coherent and fixes the center leaves, then the original diffeomorphism is itself dynamically coherent (and a good lift of a power of it will fix every center leaf).

7.1 Center leaves are all fixed

To begin, we would like to show that $\widetilde{f}$ fixes every center leaf. The results of Section 5 already provide at least one fixed center leaf:

Lemma 7.1 *Let $f: M \to M$ be an orientable partially hyperbolic diffeomorphism homotopic to the identity with f–minimal branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$. If there is a good lift $\widetilde{f}$ that fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$, then $\widetilde{f}$ fixes some center leaf.*

Proof Suppose that $\widetilde{f}$ fixes no center leaf. Since there is at least one nonplanar leaf, Proposition 5.6 provides an f–periodic center leaf c in M. Applying Proposition 5.2 to $\widetilde{\mathcal{W}}^{\mathrm{cs}}_{\mathrm{bran}}$ shows that c is coarsely contracting, but the same result applied to $\widetilde{\mathcal{W}}^{\mathrm{cu}}_{\mathrm{bran}}$ shows that c is coarsely expanding. This is a contradiction, so $\widetilde{f}$ must fix a center leaf, as desired. □

Proposition 7.2 *Let $f: M \to M$ be an orientable partially hyperbolic diffeomorphism homotopic to the identity with f–minimal branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$. If a good lift $\widetilde{f}$ of f fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$, then $\widetilde{f}$ fixes every center leaf.*

Proof Let

$$\mathrm{Fix}^{\mathrm{c}}_{\widetilde{f}} := \{c : \widetilde{f}(c) = c\},$$

thought of as a subset of the center leaf space.

The set $\mathrm{Fix}^{\mathrm{c}}_{\widetilde{f}}$ is obviously $\pi_1(M)$–invariant. It is also open, by an argument very similar to the one in [3, Lemma 6.3]: if c is a fixed center leaf in a center stable leaf L in $\widetilde{M}$, then for any center leaf c' in L close enough to c (for the topology of the center leaf space in L), there is a strong stable leaf that intersects c, c' and $\widetilde{f}(c')$. Now, since $\widetilde{f}$ fixes the center unstable leaves, c' and $\widetilde{f}(c')$ are on the same center unstable leaf. Since no transversal can intersect the same leaf twice, it implies that $c' = \widetilde{f}(c')$. Thus the set of fixed center leaves within each center stable leaf is open (in the center leaf space within that center stable leaf). Similarly, the set of fixed center leaves within each center unstable leaf is open. Together, these facts imply that the set of fixed center leaves is open in the center leaf space.

Note that a good lift $\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, so f fixes every leaf of $\mathcal{W}^{\mathrm{cs}}$. In particular, f–minimality of $\mathcal{W}^{\mathrm{cs}}$ is equivalent to minimality of $\mathcal{W}^{\mathrm{cs}}$. Hence $\mathcal{W}^{\mathrm{cs}}$ is minimal; similarly for $\mathcal{W}^{\mathrm{cu}}$.[6]

To see that $\widetilde{f}$ fixes every center leaf, we proceed as in [3, Lemma 6.4]: we show first that every center leaf in a center stable leaf (resp. center unstable leaf) which projects to an annulus has to be fixed (due to our orientability assumptions, leaves cannot project to a Möbius band). Then the same argument as in [3, Lemma 6.4] applies to show that every center leaf has to be fixed.

Let L be any center stable leaf that projects to an annulus, and choose a generator γ of the isotropy group of L.

Since the set of fixed center leaves is open in the center leaf spaces of any center unstable leaf, minimality of $\mathcal{W}^{\mathrm{cs}}$ implies that L must have some fixed center leaves.

We will first prove that if c is a center leaf in L which is in the boundary of the set of fixed center leaves in L, then $\pi(c)$ is periodic under f. We will then show, as in Proposition 5.4, that any periodic leaf in $\pi(L)$ must be coarsely contracting. The same argument applied to the center unstable leaves yields that periodic center leaves must also be coarsely expanding; a contradiction.

Since $\widetilde{f}$ cannot have fixed points (as $\widetilde{f}$ fixes all the leaves of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$), $\widetilde{f}$ acts freely on the space of stable leaves in L.

We assume, for a contradiction, that not all center leaves in L are fixed. Let Fix_L be the set (in $\mathcal{L}^{\mathrm{c}}_L$, the center leaf space on L) of center leaves fixed by $\widetilde{f}$.

[6] In fact f–minimality and minimality are always equivalent as long as the branching foliation does not have a compact leaf, without assumptions on f; see Lemma B.2.

The set Fix_L is open, and assumed not to be the whole of L. So let c_1 be any leaf in $\partial\mathrm{Fix}_L$.

Let (c_n) be any sequence of center leaves in Fix_L that converge to c_1. Then $\tilde{f}(c_n) = c_n$ converges to $\tilde{f}(c_1)$. As the leaf c_1 is not fixed by $\tilde{f}$, we deduce that $\tilde{f}(c_1)$ is not separated from c_1.

Hence, there exists a (unique) stable leaf s_1 which separates $\tilde{f}(c_1)$ from c_1 and makes a perfect fit with c_1; see Section 3.4.3 for the definition of perfect fits in the non-dynamically-coherent setting. Then $\tilde{f}(s_1)$ makes a perfect fit with $\tilde{f}(c_1)$. Because c_1 and $\tilde{f}(c_1)$ are not separated from each other, s_1 and $\tilde{f}(s_1)$ intersect a common transversal to the stable foliation. It follows that the stable axis of $\tilde{f}$ acting on L is a line. Thus, since γ commutes with $\tilde{f}$, the stable axis of γ is that same line. Moreover, both the stable leaves s_1 and $\tilde{f}(s_1)$ are in the axis of $\tilde{f}$.

Since the stable axis of $\tilde{f}$ acting on L is a line, the graph transform argument [3, Appendix H] applies and we obtain a curve $\hat{\eta}$, tangent to the center direction, which is fixed by both γ and $\tilde{f}$.

As s_1 makes a perfect fit with c_1, and s_1 intersects $\hat{\eta}$, we deduce that there exists a stable leaf s that intersects both c_1 and $\hat{\eta}$. Let $x = s \cap \hat{\eta}$ and $y = s \cap c_1$. We denote by J the segment of s between x and y.

Since $\hat{\eta}$ projects down to a closed curve $\pi(\hat{\eta})$, and $\tilde{f}$ decreases stable lengths, there exist $n_1, n_2 \in \mathbb{Z}$ and $m_1, m_2 \in \mathbb{N}$ as large as we want such that the four points $\gamma^{n_1} \tilde{f}^{m_1}(x)$, $\gamma^{n_1} \tilde{f}^{m_1}(y)$, $\gamma^{n_2} \tilde{f}^{m_2}(x)$ and $\gamma^{n_2} \tilde{f}^{m_2}(y)$ are all in a disk of radius as small as we want.

Suppose now that $\gamma^{n_1} \tilde{f}^{m_1}(c_1) \neq \gamma^{n_2} \tilde{f}^{m_2}(c_1)$. Then, up to switching n_1, m_1 and n_2, m_2, we obtain that $\gamma^{n_2} \tilde{f}^{m_2}(c_1)$ intersects $\gamma^{n_1} \tilde{f}^{m_1}(J)$. This is in contradiction with the fact that c_1 is in $\partial\mathrm{Fix}_L$, which is invariant by both $\tilde{f}$ and γ.

Thus $\gamma^{n_1} \tilde{f}^{m_1}(c_1) = \gamma^{n_2} \tilde{f}^{m_2}(c_1)$. In other words, c_1 is fixed by the map $h = \gamma^n \tilde{f}^m$ for some n, m integers with $m > 0$. (Although not useful for the rest of the proof, one can further notice that $\hat{\eta}$ and c_1 intersect, as h decreases the length of J by forward iterations and both c_1 and $\hat{\eta}$ are fixed by h.)

Now recall that we built above a stable leaf s_1 making a perfect fit with c_1. And, by our choice of s_1, the center leaf c_1 is in between s_1 and $s_2 := \tilde{f}^{-1}(s_1)$.

Recall that s_1 is the unique leaf making a perfect fit with c_1 and separating c_1 from $\tilde{f}(c_1)$. Thus $h(s_1)$ is the unique leaf making a perfect fit with $h(c_1) = c_1$ and separating

$h(c_1) = c_1$ from $h \circ \widetilde{f}(c_1) = \widetilde{f} \circ h(c_1) = \widetilde{f}(c_1)$. That is, s_1 is fixed by h. Using again that h and $\widetilde{f}$ commute, we deduce that s_2 is also fixed by h.

Now, the leaves s_1 and s_2 are also a bounded distance apart, so Lemma 5.5 holds and we deduce that c_1, as well as any other center leaf c that is in between s_1 and s_2, must be coarsely contracting. Note now that any center leaf c in L that is fixed by some $h' = \gamma^{n'} \widetilde{f}^{m'}$ is separated from Fix_L by a center leaf $c_1' \subset \partial \mathrm{Fix}_L$ as above. Hence, we proved that every nonfixed periodic leaf in $\pi(L)$ is coarsely contracting.

Therefore, the same argument applied to the center *unstable* leaf containing c_1 shows that c_1 must also be coarsely expanding; a contradiction.

So we obtained that every center stable or center unstable leaf L which is fixed by some nontrivial element of $\pi_1(M)$ has all of its center leaves fixed by $\widetilde{f}$. Since $\mathrm{Fix}^{\mathrm{c}}_{\widetilde{f}}$ is open (in the center leaf space), minimality of the foliations implies that it contains every center leaf, as in the end of the proof of [3, Lemma 6.4]. □

7.2 Dynamical coherence

We now want to prove dynamical coherence provided that a good lift fixes every center leaf. We do not assume that f is orientable, only that it admits branching foliations. We start with the following:

Lemma 7.3 *Let $f : M \to M$ be a partially hyperbolic diffeomorphism homotopic to the identity, preserving branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$. Let $\widetilde{f}$ be a good lift that fixes every center leaf. Then there is a global bound on the length of every center segment between a point x and $\widetilde{f}(x)$.*

In the dynamically coherent case this was very easy as the center curves form an actual foliation and there is a local product picture near any compact segment. We have to be more careful in the non-dynamically-coherent setting.

Proof We assume the conclusion of the lemma fails. Then there exists a sequence x_i of points in $\widetilde{M}$ contained in center leaves c_i such that the length in c_i from x_i to $\widetilde{f}(x_i)$ diverges to infinity. This length depends not only on x_i but also on c_i, since there may be many center leaves through x_i. We denote by e_i the segment in c_i from x_i to $\widetilde{f}(x_i)$.

Up to acting by covering translations we can assume that the x_i converge to a point $x \in \widetilde{M}$. Let L_i and U_i be, respectively, a center stable leaf and center unstable leaf containing c_i. Up to considering a subsequence, we may assume that L_i converges to a center stable leaf L containing x; see condition (iv) of Definition 3.1. Similarly, we can further assume that U_i converges to some center unstable leaf U with $x \in U$.

For i large enough, all the leaves L_i intersect a small unstable segment in $u(x)$. The set of center stable leaves intersecting this segment is also a segment (even though many different leaves may intersect a given point in $u(x)$). Hence we may assume that L_i is weakly monotone, and so is U_i. Let c be the center leaf through x contained in $L \cap U$. Then $\tilde{f}(x) \in c$, and we call e the segment in c from x to $\tilde{f}(x)$.

Suppose first that $L_i = L$ for all big i. So we may assume $L_i = L$ for all i. Then the center leaves c_i are all in L and, for i big enough, intersect $s(x)$. Hence the leaves c_i are, for i big enough, contained in an interval of the center leaf space in L. In addition they are converging to c, which is a center leaf through x and $\tilde{f}(x)$. This implies that the length of e_i is converging to the length of e, and hence the length of e_i is bounded in i; contradiction.

Suppose now that the L_i are all distinct from L. The points x_i and $\tilde{f}(x_i)$ are all in a compact region of $\widetilde{M}$. Since L_i converges to L, we have that $u(x_i)$ intersects L for big enough i. We call this nearby intersection y_i. Likewise, $u(\tilde{f}(x_i))$ intersects L in $\tilde{f}(y_i)$. We want to push the center segments c_i contained in $U_i \cap L_i$ along unstable segments to center segments in $U_i \cap L$.

For i big enough, both x_i and $\tilde{f}(x_i)$ are very near L. Thus, their unstable leaves $u(x_i)$ and $u(\tilde{f}(x_i))$ both intersect L. Let y_i be the intersection of $u(x_i)$ with L — recall that this intersection is unique as the center stable branching foliation is approximated by a taut foliation. Then $\tilde{f}(y_i)$ is the intersection of $u(\tilde{f}(x_i))$ with L, since L is fixed by $\tilde{f}$. Then the intersection of the unstable saturation of e_i with L is a compact segment inside a center leaf between y_i and $\tilde{f}(y_i)$, since $\tilde{f}$ fixes every center leaf. Let b_i be this segment between y_i and $\tilde{f}(y_i)$. The segments b_i also converge to e, so the previous paragraph shows that the lengths of the b_i are bounded. Since the distance between x_i and y_i converges to zero, this in turn implies that the lengths of the segments e_i are themselves bounded. This contradicts our assumption and finishes the proof. □

Lemma 7.4 *Let $f : M \to M$ be a partially hyperbolic diffeomorphism homotopic to the identity, preserving branching foliations $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$. Let $\tilde{f}$ be a good lift that fixes every center leaf. If c_1 and c_2 are different center leaves in a single center stable leaf $L \in \widetilde{\mathcal{W}}^{cs}$, then $c_1 \cap c_2 = \varnothing$.*

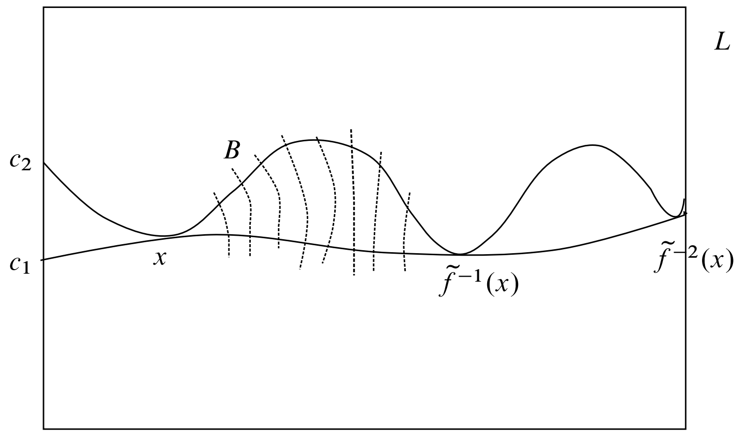

Figure 7: Two centers that merge. The bound on the distance between x and $\widetilde{f}(x)$ forces a behavior like the figure.

Proof Suppose that there are distinct center leaves c_1 and c_2 that intersect at a point $x \in c_1 \cap c_2$. Then $\widetilde{f}(x)$ is also in $c_1 \cap c_2$. If c_1 coincides with c_2 in their respective segments from x to $\widetilde{f}(x)$, then applying iterates of $\widetilde{f}$ implies that $c_1 = c_2$, contrary to assumption.

So we may assume that x is a boundary point of an open interval I in, say, c_1, which is disjoint from c_2 but such that both endpoints are in c_2. Then $c_1 \cup c_2$ bounds a bigon B with endpoints x, y and a "side" in I. All center segments in B pass through x and y and they have bounded length by Lemma 7.3. Each stable segment intersecting I also intersects the other "boundary" component of B. See Figure 7.

The stable lengths grow without bound under negative iterates of $\widetilde{f}$. Hence, since a stable segment can intersect a local foliated disk of the stable foliation in L only in a bounded length, it follows that the diameter in $\widetilde{f}^n(L)$ of $\widetilde{f}^n(B)$ grows without bound as n goes to $-\infty$. But the length of the center segments in $\widetilde{f}^n(B)$ are all bounded, according to Lemma 7.3. Moreover, between any two points in $\widetilde{f}^n(B)$ there exists a path along (at most) two center leaves — one just follows the center leaf to one of the endpoints and then switches to the appropriate other center leaf. Thus the diameter is bounded, which is a contradiction. □

Thus we deduce what we wanted to obtain in this section.

Corollary 7.5 *Let $f: M \to M$ be a partially hyperbolic diffeomorphism homotopic to the identity, preserving branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$. If some good lift $\widetilde{f}$ fixes every center leaf, then f is dynamically coherent.*

Proof By Proposition B.3 it is enough to show that the leaves of the branching foliations do not merge.

Assume that two center unstable leaves U_1 and U_2 merge. Let L be a center stable leaf intersecting U_1 and U_2 at the merging, ie L is a leaf through a point x such that the unstable leaf through x is a boundary component of $U_1 \cap U_2$. Then, connected components of $U_1 \cap L$ and $U_2 \cap L$ give two center leaves that intersect but do not coincide. This contradicts Lemma 7.4. A symmetric argument gives that two center stable leaf cannot merge either, proving dynamical coherence of f. □

7.3 Dynamical coherence without taking lifts and iterates

We now want to prove that if a finite lift and finite power of a partially hyperbolic diffeomorphism is dynamically coherent, then the original diffeomorphism is itself dynamically coherent. Although we do not know how to prove it in this generality, we show it when a good lift of the dynamically coherent lift fixes every center leaf, which is enough for our purposes.

Again, in this subsection we do not assume that f is orientable.

We start by showing a uniqueness result for the pairs of the center stable and center unstable foliations under some conditions.

Lemma 7.6 *Let $g: M \to M$ be a dynamically coherent partially hyperbolic diffeomorphism homotopic to the identity. Let $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ be g–invariant foliations tangent to E^{cs} and E^{cu}, respectively. Let $\mathcal{W}^{\mathrm{c}}$ be the center foliation associated with $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ (defined as in Definition 3.7), and assume that there exists a good lift $\widetilde{g}$ which fixes all the leaves of $\widetilde{\mathcal{W}}^{\mathrm{c}}$.*

Suppose that $\mathcal{W}_1^{\mathrm{cs}}$ and $\mathcal{W}_1^{\mathrm{cu}}$ are two g–invariant foliations tangent, respectively, to E^{cs} and E^{cu}. Suppose that $\widetilde{g}$ also fixes all the leaves of the center foliation $\widetilde{\mathcal{W}}_1^{\mathrm{c}}$ associated with $\mathcal{W}_1^{\mathrm{cs}}$ and $\mathcal{W}_1^{\mathrm{cu}}$.

Then $\mathcal{W}^{\mathrm{cs}} = \mathcal{W}_1^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}} = \mathcal{W}_1^{\mathrm{cu}}$.

Proof The argument is similar to the one made in Lemma 7.4.

Let $\widetilde{\mathcal{W}}_1^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}_1^{\mathrm{cu}}$ be two g–equivariant foliations as in the lemma. We will consider the center foliation $\widetilde{\mathcal{W}}_1^{\mathrm{c}}$ defined by taking the connected components of intersections of leaves of $\widetilde{\mathcal{W}}_1^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$, to show that $\widetilde{\mathcal{W}}^{\mathrm{cs}} = \widetilde{\mathcal{W}}_1^{\mathrm{cs}}$. A symmetric argument shows that $\widetilde{\mathcal{W}}^{\mathrm{cu}} = \widetilde{\mathcal{W}}_1^{\mathrm{cu}}$.

Since every leaf of both $\widetilde{\mathcal{W}}^{\mathrm{c}}$ and $\widetilde{\mathcal{W}}_1^{\mathrm{c}}$ is fixed by $\widetilde{g}$, Lemma 7.3 implies that $\widetilde{g}$ moves points a uniformly bounded amount in both center foliations.

Consider, for a contradiction, a point $x \in \widetilde{M}$ such that $\widetilde{\mathcal{W}}^{\mathrm{c}}(x) \neq \widetilde{\mathcal{W}}_1^{\mathrm{c}}(x)$; note that we are dealing here with actual foliations, not branching ones, so this notation makes sense. Without loss of generality, we can choose x so that the leaves $L := \widetilde{\mathcal{W}}^{\mathrm{cs}}(x)$ and $L_1 := \widetilde{\mathcal{W}}_1^{\mathrm{cs}}(x)$ do not coincide in any neighborhood of x.

Let c and c_1 be the center leaves obtained respectively as the connected components of $L \cap F$ and $L_1 \cap F$ containing x for some $F \in \widetilde{\mathcal{W}}^{\mathrm{cu}}$.

By assumption, both c and c_1 are fixed by $\widetilde{g}$, so we are in the exact same setup as in the proof of Lemma 7.4. Thus we deduce that $c = c_1$, a contradiction. □

We can now state and prove the aim of this section.

Proposition 7.7 *Let $f : M \to M$ be a partially hyperbolic diffeomorphism such that f^k is homotopic to the identity for some $k > 0$. Let $\widehat{M}$ be a finite cover of M which makes all bundles orientable. Let g be a lift to $\widehat{M}$ of a homotopy of f^k to the identity that preserves orientation of the bundles. Suppose that g is dynamically coherent and that there exists a good lift $\widetilde{g}$ of g that fixes all the center leaves. Then f is dynamically coherent and f^k is a discretized Anosov flow.*

Proof First we notice that the assumptions of the proposition will be verified for any further finite cover $\overline{M}$ of $\widehat{M}$ — because one can take a further lift $\overline{g}$ of g to $\overline{M}$, it is dynamically coherent and $\widetilde{g}$ is a good lift of $\overline{g}$ too. Hence, without loss of generality, we may and do assume that $\widehat{M}$ is a normal cover of M.

Let $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$ be the lifts to $\widetilde{M}$ of the center stable and center unstable foliations of g. Our goal is to show that these foliations are $\pi_1(M)$–invariant, thus they descend to foliations in M, and that these projected foliations are f–invariant.

Notice that $\widetilde{g}$ fixes each leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$.

The map g is obtained from a lift of a homotopy of f^k to the identity. Lifting that homotopy further to $\widetilde{M}$, we get a good lift $\widetilde{f^k}$ of f^k that is also a lift (hence a good lift) of g to $\widetilde{M}$. As both $\widetilde{g}$ and $\widetilde{f^k}$ are good lifts of g, there exists a $\beta \in \pi_1(\widehat{M}) \subset \pi_1(M)$ such that $\widetilde{g} = \beta \widetilde{f^k}$. (Note however that $\widetilde{g}$ is not necessarily a good lift of f^k as $\widetilde{g}$ only commutes with elements of $\pi_1(\widehat{M})$ and not $\pi_1(M)$.)

Moreover, both $\widetilde{g}$ and $\widetilde{f^k}$ move points a bounded distance in $\widetilde{M}$; hence so does $\beta = \widetilde{g}(\widetilde{f^k})^{-1}$. Lemma A.1 then implies that either β is the identity or M is Seifert (and β is either the identity or a power of a regular fiber).

We split the rest of the proof of dynamical coherence into two cases.

Case 1 Suppose that M is *not* a Seifert fibered space.

Then β is the identity, which means that $\widetilde{g} = \widetilde{f^k}$.

Let γ be a deck transformation in $\pi_1(M)$. Define the foliations

$$\mathcal{F}^{cs}_\gamma := \gamma\widetilde{\mathcal{W}}^{cs}, \quad \mathcal{F}^{cu}_\gamma := \gamma\widetilde{\mathcal{W}}^{cu}, \quad \mathcal{F}^{c}_\gamma := \gamma\widetilde{\mathcal{W}}^{c}.$$

The leaves of these foliations are all fixed by $\widetilde{g}$ because γ commutes with $\widetilde{f^k} = \widetilde{g}$. In particular, Lemma 7.6 then implies that $\gamma\widetilde{\mathcal{W}}^{cs} = \widetilde{\mathcal{W}}^{cs}$ and $\gamma\widetilde{\mathcal{W}}^{cu} = \widetilde{\mathcal{W}}^{cu}$. Since this is true for any element of $\pi_1(M)$, these foliations descend to foliations $\mathcal{W}^{cs}_M$ and $\mathcal{W}^{cu}_M$ in M.

Now we need to show that $\mathcal{W}^{cs}_M$ and $\mathcal{W}^{cu}_M$ are also f–invariant. Equivalently, we need to show that $\widetilde{\mathcal{W}}^{cu}$ and $\widetilde{\mathcal{W}}^{cs}$ are invariant by any lift f_1 of f to $\widetilde{M}$.

Let f_1 be a lift of f to $\widetilde{M}$. Notice that f may not be homotopic to the identity, so f_1 is not assumed to be a good lift. Let $\mathcal{F}^{cs}_1 := f_1(\widetilde{\mathcal{W}}^{cs})$ and $\mathcal{F}^{cu}_1 := f_1(\widetilde{\mathcal{W}}^{cu})$.

We will first show that f_1 and $\widetilde{g}$ commute. Both $f_1\widetilde{g}$ and $\widetilde{g} f_1$ are lifts of the map f^{k+1} to $\widetilde{M}$. So $(\widetilde{g})^{-1}(f_1)^{-1}\widetilde{g} f_1$ is a deck transformation $\gamma \in \pi_1(M)$. As $\widetilde{g}$ moves points a bounded distance, we have that $d(f_1(y), \widetilde{g} f_1(y))$ is bounded in $\widetilde{M}$. In addition, f_1 has bounded derivatives so $d(y, (f_1)^{-1}\widetilde{g} f_1(y))$ is also bounded in $\widetilde{M}$. So using again that $\widetilde{g}$ is a good lift, we deduce that $d(y, (\widetilde{g})^{-1}(f_1)^{-1}\widetilde{g} f_1(y))$ is bounded in $\widetilde{M}$.

Hence γ is a deck transformation that moves points a bounded distance. Applying Lemma A.1 again gives that β is the identity (since M is not Seifert). Hence f_1 and $\widetilde{g}$ commute.

Since $\widetilde{g}$ fixes every leaf of $\widetilde{\mathcal{W}}^{c}$ (the center foliation in $\widetilde{M}$) and commutes with f_1, we deduce that $\widetilde{g}$ fixes every leaf of $f_1(\widetilde{\mathcal{W}}^{c})$. We can again apply Lemma 7.6 to get that $f_1(\widetilde{\mathcal{W}}^{\mathrm{cs}}) = \widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $f_1(\widetilde{\mathcal{W}}^{\mathrm{cu}}) = \widetilde{\mathcal{W}}^{\mathrm{cu}}$. That is, the foliations $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$ are f_1–invariant. Since this holds for any lift of f, it implies that $\mathcal{W}^{\mathrm{cs}}_M$ and $\mathcal{W}^{\mathrm{cu}}_M$ are f–invariant. Hence f is dynamically coherent with foliations $\mathcal{W}^{\mathrm{cs}}_M$ and $\mathcal{W}^{\mathrm{cu}}_M$. This completes the proof that f is dynamically coherent when M is not Seifert fibered.

Case 2 Assume that M is Seifert fibered.

In this case, Lemma A.1 implies that $\beta = \widetilde{g}(\widetilde{f^k})^{-1}$ is either the identity or represents a power of a regular fiber of the Seifert fibration. In any case, β is in a normal subgroup of $\pi_1(M)$ isomorphic to $\mathbb{Z}$. Moreover, as proved earlier, $\beta \in \pi_1(\widehat{M})$.

Let $\gamma \in \pi_1(M)$ be any deck transformation. Consider the foliations $\mathcal{F}^{\mathrm{cs}}_\gamma := \gamma\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\mathcal{F}^{\mathrm{cu}}_\gamma := \gamma\widetilde{\mathcal{W}}^{\mathrm{cu}}$, as before.

We first claim that these foliations are $\widetilde{g}$–invariant. We show this for $\mathcal{F}^{\mathrm{cs}}_\gamma$, the other being analogous. Let $L \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$. We have

$$\widetilde{g}(\gamma L) = \beta\widetilde{f^k}(\gamma L) = \beta\gamma\widetilde{f^k}(L) = \gamma\beta^{\pm 1}\widetilde{f^k}(L).$$

Notice that both $\widetilde{f^k}$ (because it is a lift of g) and β (because it belongs to $\pi_1(\widehat{M})$ and the foliation $\mathcal{W}^{\mathrm{cs}}$ is defined in $\widehat{M}$) preserve the foliation $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. It follows that $\beta^{\pm 1}\widetilde{f^k}(L) \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$, so

$$\widetilde{g}(\gamma L) = \gamma\beta^{\pm 1}\widetilde{f^k}(L) \in \mathcal{F}^{\mathrm{cs}}_\gamma.$$

Thus $\mathcal{F}^{\mathrm{cs}}_\gamma$ is $\widetilde{g}$–invariant.

We now want to show that the foliations $\mathcal{F}^{\mathrm{cs}}_\gamma$, $\mathcal{F}^{\mathrm{cu}}_\gamma$ and $\mathcal{F}^{\mathrm{c}}_\gamma := \gamma\widetilde{\mathcal{W}}^{\mathrm{c}}$ are all leafwise fixed by $\widetilde{g}$.

Since $\widehat{M}$ was chosen to be a normal cover of M, any element $\gamma \in \pi_1(M)$ can be thought of as a diffeomorphism of $\widehat{M}$. Hence we can consider the foliation $\widehat{\mathcal{F}}^{\mathrm{cs}}_\gamma := \gamma\mathcal{W}^{\mathrm{cs}}$ in $\widehat{M}$. Note that $\widehat{\mathcal{F}}^{\mathrm{cs}}_\gamma$ is tangent to the center stable distribution $E^{\mathrm{cs}} \subset T\widehat{M}$, since γ preserves the tangent bundle decomposition, as it is defined by f in M. The argument above shows that $\widehat{\mathcal{F}}^{\mathrm{cs}}_\gamma$ is g–invariant.

Thus, we can consider g to be a dynamically coherent diffeomorphism for the pair of transverse foliations $\widehat{\mathcal{F}}^{\mathrm{cs}}_\gamma$ and $\mathcal{W}^{\mathrm{cu}}$. Moreover, g is homotopic to the identity and the good lift $\widetilde{g}$ fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$. Since $\widehat{M}$ is Seifert, mixed behavior is excluded (cf [3, Theorem 5.1]) and this implies that $\widetilde{g}$ must also fix every leaf of $\mathcal{F}^{\mathrm{cs}}_\gamma$.

The symmetric argument shows that $\mathcal{F}^{\mathrm{cu}}_\gamma$ is also fixed by $\widetilde{g}$. We can apply Proposition 6.1 to both $\widehat{\mathcal{F}}^{\mathrm{cs}}_\gamma$ and $\widehat{\mathcal{F}}^{\mathrm{cu}}_\gamma$, implying that they are g–minimal. To apply the proposition we need that g is orientable. Hence, the center foliation $\mathcal{F}^{c}_{\gamma}$ is fixed by $\widetilde{g}$, thanks to Proposition 7.2 (this also uses that g is orientable).

Since all the leaves of $\mathcal{F}^{c}_{\gamma}$ are fixed by $\widetilde{g}$, we can finally apply Lemma 7.6 to deduce that $\mathcal{F}^{cs}_{\gamma} = \widetilde{\mathcal{W}}^{cs}$ and $\mathcal{F}^{cu}_{\gamma} = \widetilde{\mathcal{W}}^{cu}$. As this is true for any γ, the foliations $\widetilde{\mathcal{W}}^{cs}$ and $\widetilde{\mathcal{W}}^{cu}$ descend to foliations $\mathcal{W}^{cs}_{M}$ and $\mathcal{W}^{cu}_{M}$ on M in this case too.

We now again have to show that $\mathcal{W}^{cs}_{M}$ and $\mathcal{W}^{cu}_{M}$ are f–invariant. The argument is the same for both foliations, so we only deal with $\mathcal{W}^{cs}_{M}$.

We start with a preliminary step. Let f_* be the automorphism of $\pi_1(M)$ induced by f. Let

$$A := \pi_1(\widehat{M}) \cap f_*(\pi_1(\widehat{M})) \cap \cdots \cap (f_*)^{k-1}(\pi_1(\widehat{M})).$$

The set A is a finite-index, normal subgroup of $\pi_1(M)$. Moreover, as f^k is homotopic to the identity, $f_*(A) = A$.

As we remarked at the beginning of the proof, we can without loss of generality prove the result for any further finite cover of $\widehat{M}$. Thus we choose, if necessary, a further cover so that $\pi_1(\widehat{M}) = A$. Since $f_*(A) = A$, the map f lifts to a homeomorphism $\widehat{f}$ of $\widehat{M}$.

As in the first case, we let f_1 be an arbitrary lift of $\widehat{f}$ to $\widetilde{M}$ and we define $\mathcal{F}^{cs}_1 := f_1(\widetilde{\mathcal{W}}^{cs})$ and $\mathcal{F}^{cu}_1 := f_1(\widetilde{\mathcal{W}}^{cu})$. (In particular, f_1 is also a lift of f.)

Note as before that both $\widetilde{g} f_1$ and $f_1 \widetilde{g}$ are lifts of f^{k+1}, and $\widetilde{g} f_1 (\widetilde{g})^{-1} (f_1)^{-1}$ is a bounded distance from the identity (because $\widetilde{g}$ is and f_1 has bounded derivatives). So $\delta := \widetilde{g} f_1 (\widetilde{g})^{-1} (f_1)^{-1}$ is an element of $\pi_1(M)$ a bounded distance from the identity. By Lemma A.1, δ represents a power of a regular fiber of the Seifert fibration, so is in the normal $\mathbb{Z}$ subgroup of $\pi_1(M)$ (note that since $\pi_1(M)$ is not virtually nilpotent, there exists a unique Seifert fibration on M; see Appendix A).

In addition, $\widetilde{g} f_1$ and $f_1 \widetilde{g}$ are also lifts of the homeomorphisms $g \widehat{f}$ and $\widehat{f} g$ in $\widehat{M}$ to $\widetilde{M}$. Hence δ is in $\pi_1(\widehat{M})$.

Using once more the arguments above, we get that $(f_1)^{-1} \delta f_1 (\delta)^{-1}$ is a bounded distance from the identity, and projects to the identity in M (and in $\widehat{M}$), hence it is a deck transformation η also contained in the $\mathbb{Z}$ normal subgroup of $\pi_1(M)$. Thus δ and η commute. Moreover, η is also in $\pi_1(\widehat{M})$.

Now we can show that $\widetilde{g}$ preserves $\mathcal{F}_1^{\mathrm{cs}}$. Let L be in $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Then

$$\widetilde{g}(f_1(L)) = \delta f_1(\widetilde{g}(L)) = \delta f_1(L) = f_1(\eta\delta(L)).$$

Here $\eta\delta(L)$ is in $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, because L is in $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\eta\delta$ is in $\pi_1(\widehat{M})$. Hence $\widetilde{f_1}(\eta\delta L)$ is in $f_1(\widetilde{\mathcal{W}}^{\mathrm{cs}})$ so $\widetilde{g}$ preserves $\mathcal{F}_1^{\mathrm{cs}}$.

What we proved implies that g preserves $\widehat{f}(\mathcal{W}^{\mathrm{cs}})$ in $\widehat{M}$. Now consider the pair of foliations $\widehat{f}(\mathcal{W}^{\mathrm{cs}})$ and $\mathcal{W}^{\mathrm{cu}}$. They are both invariant by g, so g is dynamically coherent for this particular pair of foliations, and $\widetilde{g}$ fixes the leaves of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$. So once again, as $\widehat{M}$ is Seifert, we get that $\widetilde{g}$ must also fix every leaf of $f_1(\widetilde{\mathcal{W}}^{\mathrm{cs}})$; cf [3, Theorem 5.1].

The symmetric argument implies that $\widetilde{g}$ fixes every leaf of $f_1(\widetilde{\mathcal{W}}^{\mathrm{cu}})$. Once again, $\widehat{M}$ being Seifert implies that all the foliations are g–minimal (Proposition 6.1). Hence $\widetilde{g}$ also fixes the center foliation $f_1(\widetilde{\mathcal{W}}^{\mathrm{c}})$ (Proposition 7.2). So Lemma 7.6 applies and we deduce that $f_1(\widetilde{\mathcal{W}}^{\mathrm{cs}}) = \widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $f_1(\widetilde{\mathcal{W}}^{\mathrm{cu}}) = \widetilde{\mathcal{W}}^{\mathrm{cu}}$.

In particular, f preserves the foliations $\mathcal{W}_M^{\mathrm{cs}}$ and $\mathcal{W}_M^{\mathrm{cu}}$ as wanted. So f is dynamically coherent.

This finishes the proof that f is dynamically coherent. Once that is known, then Propositions 6.5 and G.2 of [3] imply that f^k is a discretized Anosov flow. This finishes the proof of the proposition. □

8 Proof of Theorem A

Fix a partially hyperbolic diffeomorphism $f: M \to M$ that is homotopic to the identity on a closed Seifert fibered 3–manifold M. We make no orientability assumptions. We will show that some iterate of f is a discretized Anosov flow, completing the proof of Theorem A.

Fix a finite cover $\widehat{M}$ of M so that the lifted center, stable and unstable bundles are orientable. Then there is an integer $k > 0$ such that a lift of f^k to $\widehat{M}$ will preserve the orientations of the bundles. In addition, we can find such a lift that is homotopic to the identity by lifting a homotopy from f^k to the identity. Fix such a lift $g: \widehat{M} \to \widehat{M}$.

Applying Theorem 3.6, we have g–invariant center stable and center unstable branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ on $\widehat{M}$.

Lemma 8.1 *There exists a lift $\widetilde{g}$ of an iterate of g that fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and also fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$.*

Proof We will use the following result, found in [3, Proposition 7.1 and Remark 7.2].

Proposition 8.2 *Let $g\colon M \to M$ be a partially hyperbolic diffeomorphism that is homotopic to the identity on a Seifert fibered 3–manifold M with orientable Seifert fibration. Then some iterate of g has a good lift which fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$.*

Since $\widehat{M}$ is orientable, the bundles are orientable, and $\mathcal{W}^{\mathrm{cs}}$ is a horizontal foliation (see [3, Theorem F.3]), it follows that the Seifert fibration is orientable. Thus there is an integer $i > 0$ such that the iterate g^i has a good lift $\widetilde{g^i}$ which fixes every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$.

Suppose that $\widetilde{g^i}$ fixes one leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$. Then Proposition 6.1 says that $\mathcal{W}^{\mathrm{cu}}$ is g^i–minimal and $\widetilde{g^i}$ fixes *every* leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$, as desired.

Suppose, then, that $\widetilde{g^i}$ fixes no leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$. Then $\widetilde{g}$ fixes no center leaf, and we can apply Proposition 5.2 to see that every periodic center leaf of g has to be coarsely contracting. Exchanging roles, and applying Proposition 8.2 to the center unstable branching foliation, we deduce that every periodic center leaf for g must be coarsely expanding. Notice that although the lifts may be different, the coarsely expanding and coarsely contracting behavior is for periodic center leaves of the original map g.

As there must be at least one such periodic center leaf (cf Proposition 5.6), this gives a contradiction. □

Let $\widetilde{g^i}$ be a good lift of an iterate g^i, for some $i > 0$, that fixes every leaf of both $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ and $\widetilde{\mathcal{W}}^{\mathrm{cu}}$. Then Proposition 7.2 implies that $\widetilde{g^i}$ fixes every center leaf, and Corollary 7.5 says that g^i is dynamically coherent. Then Proposition 7.7 tells us that f is dynamically coherent.

Now that we have reduced to the dynamically coherent case, [3, Theorem A] says that f has an iterate that is a discretized Anosov flow. This completes the proof of Theorem A.

Note that the arguments in the proof of Lemma 8.1 also eliminate mixed behavior for good lifts in Seifert fibered manifolds.

9 Absolutely partially hyperbolic diffeomorphisms

In this section, we explain how one can improve the trichotomy in Section 2.0.1 eliminating the mixed case, if one uses a strong version of partial hyperbolicity.

Definition 9.1 A partially hyperbolic diffeomorphism $f: M \to M$ on a 3–manifold is called *absolutely partially hyperbolic* if there exist constants $\lambda_1 < 1 < \lambda_2$ such that for some $\ell > 0$ and every $x \in M$, we have

$$\|Df^{\ell}|_{E^{\mathrm{s}}(x)}\| < \lambda_1 < \|Df^{\ell}|_{E^{\mathrm{c}}(x)}\| < \lambda_2 < \|Df^{\ell}|_{E^{\mathrm{u}}(x)}\|.$$

Notice that, although subtle, the difference between being absolutely partially hyperbolic versus just partially hyperbolic is far from trivial. Here, we just show that with this stronger property one can significantly simplify the arguments. However, some previous results have shown significant differences between the two notions, specifically with regard to the integrability of the bundles; see [9; 31; 29].

We will show the following:

Theorem 9.2 *Let $f: M \to M$ be an absolutely partially hyperbolic diffeomorphism on a 3–manifold. Suppose that f is homotopic to the identity and preserves two branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ that are both f–minimal. Then either*

(i) *f is a discretized Anosov flow, or*

(ii) *$\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$ are $\mathbb{R}$–covered and uniform and a good lift $\widetilde{f}$ of f acts as a translation on their leaf spaces.*

In order to prove this theorem, the main step will be to show that, using absolute partial hyperbolicity, we have an improvement of Proposition 5.2.

Proposition 9.3 *Let $f: M \to M$ be an absolutely partially hyperbolic diffeomorphism homotopic to the identity, and $\widetilde{f}$ a good lift of f to $\widetilde{M}$. Assume that every leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ is fixed by $\widetilde{f}$. Let L be a leaf whose stabilizer is generated by $\gamma \in \pi_1(M) \setminus \{\mathrm{id}\}$. Then there is a center leaf in L fixed by $\widetilde{f}$.*

The proof is essentially the same as the one in [25, Section 5.4], but we repeat it since the contexts are different.

Proof The proof is by contradiction. Assume that $\widetilde{f}$ does not fix any center leaf in L.

Proposition 5.6 gives that there exists a center leaf which is periodic by f. Call c a lift of this center leaf. Using Proposition 5.7 we get two stable leaves s_1 and s_2 in L fixed by $h := \gamma^n \circ \widetilde{f}^m$, a bounded distance apart in L and such that c separates s_1 from s_2 in L. We denote by B the band bounded by s_1 and s_2.

Since γ is an isometry, the diffeomorphism h is absolutely partially hyperbolic, and we can (modulo taking iterates) assume that there are constants $\lambda_1 < \lambda_2$ such that

$$\|Dh|_{E^{\mathrm{s}}}\| < \lambda_1 < \lambda_2 < \|Dh|_{E^{\mathrm{c}}}\|.$$

Moreover, there is a constant $R > 1$ such that $\|Dh^{-1}\| \le R$ in all of L.

For simplicity, we will assume that the distance between s_1 and s_2 is smaller than $\frac{1}{2}$ so that the band B is contained in the neighborhood $\widehat{B} = \bigcup_{x \in S_1} B_1(x)$ of radius 1 around s_1.

For every positive d there is a constant $r(d) > 0$ such that for any set of diameter less than d, the length of a stable leaf contained in this set is at most $r(d)$. This is because in a foliated box only one segment of a stable segment can intersect it. This implies that stable leaves (and center leaves as well) are quasi-isometrically embedded in their neighborhoods of a fixed diameter. So there is a $K > 0$ such that for any stable segment J contained in $\widehat{B}$ with endpoints z and w, we have

$$\operatorname{length}(J) \le K d_{\widehat{B}}(z, w).$$

Now, choose $n > 0$ such that $K^2 \lambda_1^n / \lambda_2^n \ll \frac{1}{2}$ and once n is fixed, choose $D > 0$ so that $D/2 \gg 2R^n + 2K/\lambda_2^n$.

We now pick points $z, w \in s_1$ such that $d_{\widehat{B}}(z, w) = D$, and take J^{s} an arc of s_1 joining these points. From the choice of K and D we know that $\operatorname{length}(J^{\mathrm{s}}) \le KD$. So it follows that $\operatorname{length}(h^n(J^{\mathrm{s}})) \le KD\lambda_1^n$.

Choose a center curve J^{c} joining $B_1(h^n(z))$ with $B_1(h^n(w))$ — this can be done because c separates s_1 from s_2 — and call z_n and w_n the endpoints in each ball. It follows that $\operatorname{length}(J^{\mathrm{c}}) \le K^2 D\lambda_1^n + 2K$.

Since the distance between the endpoints of J^{c} and $h^n(z)$, $h^n(w)$ is less than 1, by iterating backwards by h^{-n} we get that $d(h^{-n}(z_n), z)$ and $d(h^{-n}(w_n), w)$ are less than R^n.

This implies that

$$D \le d_{\widehat{B}}(z, w) \le K^2 \frac{\lambda_1^n}{\lambda_2^n} D + 2R^n + \frac{2K}{\lambda_2^n},$$

a contradiction with the choices of n and D, completing the proof of the proposition. □

Using this proposition, we can prove Theorem 9.2 in the same way as [3, Theorem 5.1].

Proof of Theorem 9.2 Let $\tilde{f}$ be a good lift of f. Since $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$ are f–minimal, by Corollary 4.9 $\tilde{f}$ either fixes each leaf of $\widetilde{\mathcal{W}}^{cs}$ and $\widetilde{\mathcal{W}}^{cu}$, or acts as a translation on both leaf spaces (in which case the foliations are $\mathbb{R}$–covered and uniform and we are in case (ii) of the theorem), or $\tilde{f}$ translates one and fixes the other.

If $\tilde{f}$ fixes the leaves of both $\widetilde{\mathcal{W}}^{cs}$ and $\widetilde{\mathcal{W}}^{cu}$, then Proposition 7.2 and Corollary 7.5 imply that we are in case (i) of the theorem.

So we have to show that we cannot be in the mixed case. Suppose that $\tilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{cs}$.

Since M is not $\mathbb{T}^3$, there are leaves of $\mathcal{W}^{cs}$ with nontrivial fundamental group. Consider the lift L in $\widetilde{\mathcal{W}}^{cs}$ of such a leaf, with L invariant by γ in $\pi_1(M) \setminus \{\mathrm{Id}\}$. We can apply Proposition 9.3 to conclude that there is a center leaf c in L that is fixed by $\tilde{f}$. So, in particular, $\tilde{f}$ needs to fix a center unstable leaf containing c (note that there may be an interval of center unstable leaves intersecting L in c, but the endpoints of such an interval will then be fixed by $\tilde{f}$). Thus $\tilde{f}$ has to also fix every leaf of $\widetilde{\mathcal{W}}^{cu}$, by Corollary 4.9. □

10 Regulating pseudo-Anosov flows and translations

The rest of the paper is concerned with hyperbolic 3–manifolds. We will get positive results dealing with the non-dynamically-coherent case. That is, we want to understand the dynamics of a homeomorphism acting by translation on a branching foliation. In order to be able to do that, we first need to build a regulating pseudo-Anosov flow transverse to the branching foliation. The existence of such a flow is a relatively immediate consequence of the construction of the regulating flow and the fact that the branching foliation is well-approximated by foliations.

Proposition 10.1 *Let M be a hyperbolic 3–manifold and $\mathcal{F}$ a branching foliation well-approximated by foliations $\mathcal{F}_\epsilon$ and such that $\mathcal{F}$ (and thus also $\mathcal{F}_\epsilon$ for small ϵ) is $\mathbb{R}$–covered and uniform. Then there exists a transverse and regulating pseudo-Anosov flow Φ for $\mathcal{F}$.*

Proof By [33; 11; 17] (see [3, Theorem D.3]), for any ϵ there exists a pseudo-Anosov flow Φ_ϵ transverse to and regulating for $\mathcal{F}_\epsilon$.

Now, as ϵ gets small, the angle between leaves of $\mathcal{F}_\epsilon$ and leaves of $\mathcal{F}$ becomes arbitrarily small.

Then, since both $\mathcal{F}$ and $\mathcal{F}_\epsilon$ are $\mathbb{R}$–covered and uniform, for any leaf $L \in \widetilde{\mathcal{F}}$ there exist two leaves $L_1, L_2 \in \widetilde{\mathcal{F}}_\epsilon$ such that L is in between L_1 and L_2.[7] As Φ_ϵ is regulating for $\mathcal{F}_\epsilon$, every orbit of $\widetilde{\Phi}_\epsilon$ intersects both L_1 and L_2, thus it also intersects L. So every orbit of $\widetilde{\Phi}_\epsilon$ intersects every leaf of $\widetilde{\mathcal{F}}$; that is, Φ_ϵ is regulating for $\mathcal{F}$.

The fact that the flow Φ_ϵ can be chosen transverse to $\widetilde{\mathcal{F}}$ follows from the construction of Φ_ϵ; see [33; 11; 17]. The flow Φ_ϵ is build by blowing down certain laminations transverse to $\mathcal{F}_\epsilon$. Moreover, these laminations are transverse to any foliations that are close enough to $\mathcal{F}_\epsilon$ for a uniform angle. Since the angle between $\mathcal{F}$ and $\mathcal{F}_\epsilon$ gets arbitrarily small, Φ_ϵ will also be transverse. For a continuous family of $\mathbb{R}$–covered foliations, this property is explicitly stated in [11, Corollary 5.3.22]. □

Using the regulating pseudo-Anosov flow given by Proposition 10.1, all of [3, Section 8] works for a branching foliation without change. Thus we obtain:

Proposition 10.2 *Let M be a hyperbolic 3–manifold. Let $f : M \to M$ be a homeomorphism homotopic to the identity that preserves a (branching) foliation $\mathcal{F}$. Suppose that $\mathcal{F}$ is uniform and $\mathbb{R}$–covered, and that a good lift $\widetilde{f}$ of f acts as a translation on the leaf space of $\mathcal{F}$. Let Φ be a transverse regulating pseudo-Anosov flow to $\mathcal{F}$.*

Then, for every $\gamma \in \pi_1(M)$ associated with a periodic orbit of Φ, there is a compact $\widehat{f}_\gamma$–invariant set T_γ in M_γ which intersects every leaf of $\widehat{\mathcal{F}}_\gamma$, where $M_\gamma = \widetilde{M}/\langle\gamma\rangle$ and $\widehat{f}_\gamma : M_\gamma \to M_\gamma$ is the corresponding lift of f.

Moreover, if an iterate $\widehat{f}_\gamma^k$ of $\widehat{f}_\gamma$ fixes a leaf L of $\widehat{\mathcal{F}}_\gamma$, and γ fixes all the prongs of this orbit, then the fixed set of $\widehat{f}_\gamma^k$ in L is contained in $T_\gamma \cap L$ and has negative Lefschetz index.

Almost without any change, we obtain the corresponding version of [3, Proposition 9.1].

Proposition 10.3 *Let f be a partially hyperbolic diffeomorphism in a hyperbolic 3–manifold which preserves a branching foliation $\mathcal{W}^{cs}$ tangent to E^{cs}. Assume that a good lift $\widetilde{f}$ of f acts as a translation on the foliation $\mathcal{W}^{cs}$, and let Φ^{cs} be a transverse regulating pseudo-Anosov flow. Then, for every $\gamma \in \pi_1(M)$ associated to the inverse periodic orbit of Φ^{cs}, there are $n > 0$ and $m > 0$ such that $h = \gamma^n \circ \widetilde{f}^m$ fixes a leaf L of $\mathcal{W}^{cs}$.*

[7] By construction, each leaf of $\mathcal{F}$ is the image of a leaf of $\mathcal{F}_\epsilon$ by a continuous map homotopic to the identity of M, so, given a leaf $L \in \widetilde{\mathcal{F}}$, there is a leaf $L' \in \widetilde{\mathcal{F}}_\epsilon$ at a bounded distance $< a_1$ from L. Now using the fact that $\mathcal{F}_\epsilon$ is uniform, choose L_1, L_2 in $\widetilde{\mathcal{F}}_\epsilon$ on different components of $\widetilde{M} - L'$, and so that for any $p \in L'$, $q \in L_1$ and $z \in L_2$, we have $d(p,q) > a_1$ and $d(p,z) > a_1$. The leaves L_1 and L_2 satisfy the required property.

Proof The only difference is that we cannot say that the action of h in the leaf space is expanding, since collapsing of leaves may change the behavior. However, the same proof gives the existence of an interval in the leaf space which is mapped inside itself by h^{-1} giving a fixed leaf, as desired. $\square$

Remark 10.4 In the non-dynamically-coherent situation, the proof of [3, Theorem B] does not give a contradiction: it could happen (and indeed does happen in a situation with similar properties, see eg [7]) that having a fixed point in a leaf of the foliation does not force the dynamics on the leaf space to be repelling around the leaf in terms of the action on the leaf space. This issue has previously appeared, in particular in Proposition 6.2.

Notice that if one assumes the existence of a periodic center leaf, then we can easily prove a version of [3, Theorem B] in the non-dynamically-coherent setting.

Proposition 10.5 *Let $f\colon M \to M$ be a partially hyperbolic diffeomorphism on a hyperbolic 3–manifold. Suppose that there exists a closed center leaf c that is periodic under f. Then f is a discretized Anosov flow.*

Proof We start by replacing f by a power, so that f becomes homotopic to the identity.

Let $\widetilde{f}$ be a good lift of f. We will show that $\widetilde{f}$ fixes every leaf of $\widetilde{\mathcal{W}}^{cs}$ and $\widetilde{\mathcal{W}}^{cu}$. Then Section 7 above shows that the original f (before taking a power) is dynamically coherent; hence the result follows from [3, Theorem B].

Suppose that $\widetilde{f}$ does not fix every leaf of, say, $\widetilde{\mathcal{W}}^{cs}$. Then Corollary 4.9 implies that the leaf space of $\widetilde{\mathcal{W}}^{cs}$ is $\mathbb{R}$ and that $\widetilde{f}$ acts as a translation on it.

Let $\widetilde{c}$ be a lift of the periodic closed center leaf c. Since c is periodic and $\widetilde{f}$ acts as a translation, there exists $\gamma \in \pi_1(M)$ which is nontrivial and such that $\gamma(\widetilde{c}) = \widetilde{f}^k(\widetilde{c})$ for some k. Now c is also closed, so there exists $g \in \pi_1(M) - \mathrm{Id}$ such that $g(\widetilde{c}) = \widetilde{c}$. We have that g is distinct from any power of γ, since if $L \in \widetilde{\mathcal{W}}^{cs}$ is such that $\widetilde{c} \in L$, we have that $g(L) = L \neq \gamma^k(L)$ for every $k \neq 0$.

On the other hand, $g\circ\gamma(\widetilde{c}) = g\circ\widetilde{f}^k(\widetilde{c}) = \widetilde{f}^k\circ g(\widetilde{c}) = \gamma(\widetilde{c})$, which implies that $\gamma^{-1}\circ g\circ\gamma$ and g fix $\widetilde{c}$. This is impossible since M is hyperbolic: if they both fix $\widetilde{c}$ then they have they have the same axis. But the geodesic axes of the hyperbolic transformations g and $\gamma^{-1}g\gamma$ cannot share an ideal point since g and γ are not contained in a cyclic group. $\square$

Remark 10.6 The arguments here show that the dynamics of the transverse pseudo-Anosov flow coarsely affects the dynamics of f. In particular, if $\widetilde{f}$ is a translation with respect to a certain $\mathbb{R}$–covered branching foliation, there must be a lower bound on the topological entropy of f depending only on the $\mathbb{R}$–covered branching foliation and the amount of translation of $\widetilde{f}$. It is possible that in certain hyperbolic 3–manifolds one could control the possible geometries of $\mathbb{R}$–covered foliations, in which case one could find a uniform lower bound on the entropy of partially hyperbolic diffeomorphisms that act as translations on their branching foliations. If such a bound could be obtained, one could deduce that if the entropy of a partially hyperbolic diffeomorphism is sufficiently low, then the system must be a *discretized Anosov flow*.

11 Translations in hyperbolic 3–manifolds

In this section we obtain further consequences of having a partially hyperbolic diffeomorphism act as a translation in a hyperbolic 3–manifold.

We start by recalling the setting. Let $f: M \to M$ be a (not necessarily dynamically coherent) partially hyperbolic diffeomorphism on a hyperbolic 3–manifold. Up to replacing f by a power, we assume that it is homotopic to the identity. Up to taking a further iterate of f and a lift to a finite cover of M, we can assume that f admits branching foliations, and that the good lift $\widetilde{f}$ acts as a translation on the leaf space of $\widetilde{\mathcal{W}}^{cs}$.

Let Φ^{cs} be a transverse regulating pseudo-Anosov flow to $\mathcal{W}^{cs}$ given by Proposition 10.1. This flow is fixed throughout the discussion.

Then Proposition 10.3 shows that for any periodic orbit of Φ^{cs}, there exists a center stable leaf periodic by f.

11.1 Periodic center rays

We will now produce rays in periodic center leaves which are expanding. A *ray* in L is a proper embedding of $[0, \infty)$ into L. We say that a ray is a *center ray* if it is contained in a center leaf. So a center ray c_x is the closure in L of a connected component of $c \setminus \{x\}$, where c is a center curve and $x \in c$.

Let γ in $\pi_1(M)$ be associated with a periodic orbit δ_0 of the pseudo-Anosov flow Φ^{cs}. Let L be a leaf (given by Proposition 10.3) of $\widetilde{\mathcal{W}}^{cs}$ fixed by $h := \gamma^n \circ \widetilde{f}^m$, with $m > 0$.

A center ray c_x is *expanding* if $h(c_x) = c_x$ and x is the unique fixed point of h in c_x and every $y \in c_x \setminus \{x\}$ verifies that $h^{-n}(y) \to x$ as $n \to +\infty$. It is *contracting* if it is expanding for h^{-1}.

Proposition 11.1 *Assume that a good lift $\widetilde{f}$ of f acts as a translation on the (branching) foliation $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Let Φ^{cs} be a regulating transverse pseudo-Anosov flow. Let γ in $\pi_1(M)$ be associated with a periodic orbit δ_0 of Φ^{cs}. Let L be a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ fixed by $h = \gamma^n \circ \widetilde{f}^m$, where $m > 0$. Assume that γ fixes all prongs of a lift of δ_0 to $\widetilde{M}$. Then there are at least two center rays in L, fixed by h, which are expanding.*

Remark 11.2 We should stress that we cannot guarantee that we get a single center leaf with both rays expanding. For example, it is very easy to construct an example such that h has Lefschetz index -1 in L, and has exactly 3 fixed center leaves in L, and only two fixed expanding rays, which are contained in distinct center leaves; see Figure 9. This situation occurs in the examples constructed in [7] in the unit tangent bundle of a surface.

We will use Proposition 11.1 and its proof to eliminate the mixed behavior in hyperbolic 3–manifolds. It should be noted that this proposition also gives some relevant information about the structure of the enigmatic double translation examples which are not ruled out by our study.

The key point is to understand how each fixed center leaf contributes to the total Lefschetz index of the map in a center stable leaf which we can control. Since the dynamics preserves foliations and one of them has a well-understood dynamical behavior (ie in the center stable foliation, the stable foliation is contracting) we can compute the index just by looking at the dynamics in the center foliation; see Figure 8.

As remarked above, one does have to be careful when computing the index, as cancellations might happen with branching foliation; see Figure 9.

We are now ready to give a proof of Proposition 11.1.

Proof of Proposition 11.1 By Proposition 10.2, we know that the fixed-point set of h in L is contained in the lift of T_γ to $\widetilde{M}$ (which intersects L in a compact set) and has Lefschetz index $1 - p$, where p is the number of stable prongs at the fixed point. In particular, h has some fixed points in L.

Let $L_2 = \widetilde{f}^m(L)$. We denote by $\tau_{12}\colon L \to L_2$ the flow along the $\widetilde{\Phi}^{\mathrm{cs}}$ map.

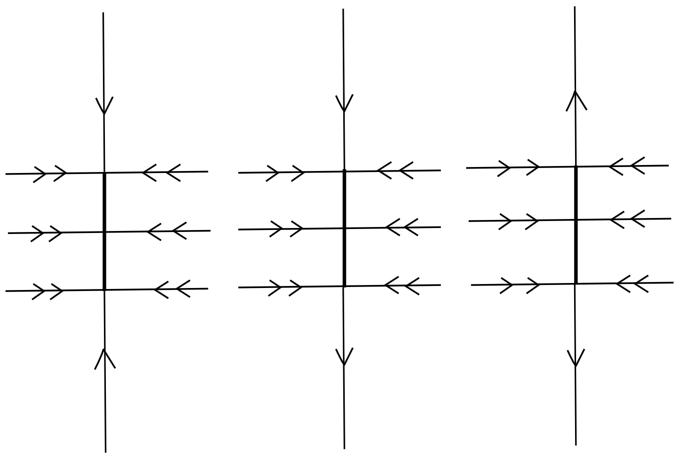

Figure 8: Contribution of index of a center arc, depending on the center dynamics.

Claim 11.3 *Let c_1 and c_2 be two distinct center leaves in L that have a nontrivial intersection. Suppose that both c_1 and c_2 are fixed by h, and there exist two distinct points $z, y \in c_1 \cap c_2$ which are fixed by h. Then the center leaves c_1 and c_2 coincide on the segment between z and y.*

Proof Let $[y, z]_{c_1}$ and $[y, z]_{c_2}$ be the center segments between y and z in c_1 and c_2, respectively.

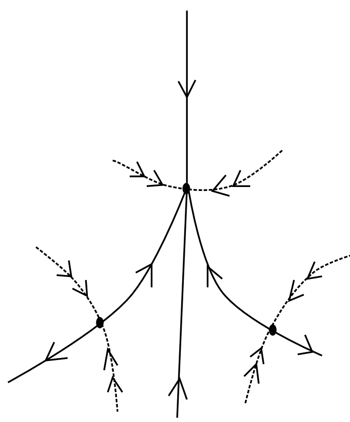

Figure 9: Two segments of zero index merge with a point with index 1 to produce a global -1 index.

Assume for a contradiction that $[y,z]_{c_1}$ and $[y,z]_{c_2}$ are distinct. Then, up to changing y and z, we can assume that the intersection between the open intervals $(y,z)_{c_1}$ and $(y,z)_{c_2}$ is empty.

Thus, by construction, $[y,z]_{c_1}$ and $[y,z]_{c_2}$ intersect only at z and y. We let B be the bigon in L bounded by $[y,z]_{c_1}$ and $[y,z]_{c_2}$.

Note that any stable leaf that enters the bigon B must exit it (otherwise it would limit in a stable leaf entirely contained in B, which is impossible). Hence, B is "product foliated" by stable leaves. Since B is compact, the length of the stable segments contained in B is bounded.

Since z, y are fixed by h it follows that B is also fixed by h. Let s be one such stable segment connecting $(z,y)_{c_1}$ to $(z,y)_{c_2}$. Then the images of s under powers of h^{-1} stay in B but must also have unbounded length, a contradiction. □

Let x be a fixed point of h. Recall from Lemma 3.19 that the set of center leaves through x in L is a closed interval. In particular, h fixes the endpoints of this interval. Hence, x is contained in a center leaf c such that $h(c) = c$.

Claim 11.4 *All the fixed points of h in L are contained in the union of finitely many compact segments of center leaves in L.*

Proof Let c be a center leaf fixed by h. Since the fixed points are contained in a compact set C (see [3, Lemma 8.11]), there is a minimal compact interval J in c which contains all the fixed points of h in c.

Suppose that there exist infinitely many distinct such minimal intervals J_i in center leaves c_i. Since the fixed points of h in L are in a compact set, we can choose i and j large enough that J_i is very close in the Hausdorff distance of L to J_j. Let z be an endpoint of J_i. Then the stable leaf $s(z)$ through z intersects the center leaf c_j. As z is fixed by h and so is c_j, contraction of the stable length implies that $z \in c_j$, thus $z \in J_j$.

Hence, both endpoints of J_i are on J_j. By Claim 11.3, it implies that $J_i \subset J_j$, and minimality of the interval J_j implies $J_j = J_i$, which is a contradiction. □

Let $\{J_i, 1 \le i \le i_0\}$ be a finite family of compact intervals containing all the fixed point of h, as given by Claim 11.4. Note that we do not necessarily take the minimal intervals as constructed in the proof of Claim 11.4, as we want the following properties for that family.

Claim 11.5 *We can choose the collection of intervals* $\{J_i, 1 \le i \le i_0\}$, *each in a center leaf fixed by* h, *satisfying the following properties*:

(1) *The union* $\bigcup_{1\le i\le i_0} J_i$ *contains all the fixed points of* h.

(2) *The endpoints of each interval* J_i *are fixed by* h.

(3) *The intervals are pairwise disjoint.*

Proof Let $c_1, \dots, c_n$ be a minimal collection of center leaves that contains all fixed points of h in L, as given by Claim 11.4. Let J_i be the minimal compact interval containing all fixed points of h in c_i.

The family J_i then satisfies conditions (1) and (2). So we only have to show that one can split the intervals J_i further so that condition (3) is also satisfied (while still satisfying the first two conditions).

Notice that c_i and c_j intersect if and only if J_i and J_j intersect. Thus, we can restrict our attention to each connected component of the union of the c_i separately.

Up to renaming, assume that $\bigcup_{1\le i\le k} c_k$ is a connected component of $\bigcup_{1\le i\le n} c_k$.

Now we can consider the union of the $J_1, \dots, J_k$ as a graph, where the vertices are the endpoints of the segments J_i together with the points where two segments merge, and the edges are the subsegments joining the vertices. With this convention, the union of the $J_1, \dots, J_k$ is then a tree. Otherwise there would be a bigon in L enclosed by the union, which is ruled out by Claim 11.3.

Let $\mathcal{B}$ be this tree. Our goal is to remove enough open segments from the J_i so that no vertex of this associated tree has degree 3 or more. Consider a vertex p in $\mathcal{B}$ with degree 3 or more. Then there are two edges e_1 and e_2 abutting at p on the same side of p. We claim that e_1 cannot have points fixed by h arbitrarily close to p (except for p itself). Otherwise one would have a fixed point $y \in e_1$ such that $s(y)$ intersects e_2. Since e_2 is contained in a fixed leaf, $e_2 \cap s(y)$ is fixed by h. This implies (since h decreases stable length) that y is in e_2. Thus, by Claim 11.3, the intersection of e_1 and e_2 would contain the segment $[y, p]$, contradicting the fact that they are distinct edges.

Thus, we can remove an open interval (p, z) from, say, e_1, where z is fixed by h but (p, z) has no fixed points. In the new tree, p has index one less than before and z has index one.

Doing this recursively on each vertex of index strictly greater than 2, we will obtain, as sought, a disjoint collection of intervals that also satisfy conditions (1) and (2). □

Now we will look at the index of h on the fixed intervals J_i, for $1 \le i \le i_0$, produced by Claim 11.5. Note that for each such interval J_i there are no other fixed points of h nearby in L. Let c be a leaf fixed by h containing J_i.

If h is contracting on c near both endpoints of J_i on the outside, then the index of J_i is $+1$. This is because the stable foliation is contracting under $h = \gamma^n \circ \widetilde{f}^m$ (since $m > 0$). Hence h is contracting near J_i. If h is expanding on both sides, the index is -1. If one side is contracting and the other is expanding, then the index is zero.

The global index for h can then be computed by adding the indexes of h on each of the intervals J_i, taking care of cancellations.

Let c_k, for $1 \le k \le k_0$, be finitely many center leaves, fixed by h and containing all the J_i. We choose this collection to have the minimum possible number of leaves.

Each leaf c_k contains finitely many segments J_i, so there are exactly two infinite rays that do not contain any J_i. The contribution of c_k to the global index of h (before possible cancellations) will then be -1 if both rays are expanding, 0 if one is expanding while the other contracts, and 1 if both are contracting.

Suppose, for a contradiction, that there is at most one expanding ray in L. So each c_k, considered separately, has index either 0 or 1.

If there is an expanding ray, let c_k be a leaf with an expanding ray. Otherwise let c_k be any leaf. Now we need to consider how the other leaves and the possible cancellations impact the global index of h. Let c_l be a leaf that intersects c_k. If c_l shares an expanding ray with c_k, then the other ray of c_l is contracting, and eventually disjoint from the corresponding ray of c_k. The fixed set (if any) of this ray in c_l has index zero. If c_l does not share an expanding ray with c_k, then both rays of c_l are contracting. The ray that is added to the same end as the expanding ray of c_k contributes index 1. The other ray contributes index 0. In any case the index, starting at 0 or 1, does not decrease.

Now, if c_m is another leaf that is disjoint from the set above, then both rays are contracting and it contributes an index 1. So again the index does not decrease.

Thus, if there is at most one expanding ray, then the index of h is at least 0. This contradicts the fact that the index of h is $1 - p$ where $p \ge 2$, and thus finishes the proof of Proposition 11.1. □

11.2 Periodic rays and boundary dynamics

Proposition 11.1 gave the existence of periodic rays that are coarsely expanding. Here we will show that such a ray has a well-defined ideal point on the circle at infinity of the leaf, and that it corresponds to the endpoint of a prong of the transverse regulating pseudo-Anosov flow, Φ^{cs}.

As previously, we assume that we have a center stable leaf $L \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$ such that there is a deck transformation γ for which $\gamma \circ \widetilde{f}^m(L) = L$ for some $m > 0$. We let $L_2 = \widetilde{f}^m(L)$ and define $\tau_{12}\colon L \to L_2$ the flow along $\widetilde{\Phi}^{\mathrm{cs}}$ map. We also take as before

$$h := \gamma \circ \widetilde{f}^m \quad \text{and} \quad g := \gamma \circ \tau_{12}.$$

Recall that h and g are maps of L that are a bounded distance from each other. Also g preserves the (singular) foliations $\mathcal{G}^{\mathrm{s}}$ and $\mathcal{G}^{\mathrm{u}}$. We again assume that if g has a fixed point x_0 in L then γ is such that g preserves each of the prongs of $\mathcal{G}^{\mathrm{s}}(x_0)$ and $\mathcal{G}^{\mathrm{u}}(x_0)$.

The action of g on the circle at infinity $S^1(L_1)$ has an even number of fixed points, which are alternately attracting and repelling. We denote by P the set of attracting fixed points and by N the set of repelling ones. With this notation, we get the following.

Proposition 11.6 *Let $\eta\colon [0,\infty) \to L$ be a contracting fixed ray for h. Then the limit $\lim_{t\to\infty} \eta(t)$ exists in $S^1(L)$ and it is a (unique) point in N. (Symmetrically, if η is an expanding fixed ray, its limit point belongs to P.)*

Proof Let y be in P and let U be a small neighborhood of y in $L \cup S^1(L)$ as in [3, Section 8]. If η has a point q in $U \cap L$, then $h^n(q)$ converges to y as $n \to +\infty$, so η could not be a contracting ray; a contradiction. So η cannot limit to any point in P. If z is in $S^1(L) \setminus \{N \cup P\}$, then $h^n(z)$ converges to a point in P under forward iteration. Hence, again, a small neighborhood Z of z in $L \cup S^1(L)$ is sent, under some iterate, inside a neighborhood U as in the first part of the proof. So any point in $Z \cap L$ converges to a point in P under forward iteration. Hence η cannot limit to a point in $S^1(L) \setminus \{N \cup P\}$ either. So η can only limit to points in N. Since η is properly embedded in L, the set of accumulations points of η is connected, so it has to be a single point. $\square$

12 Mixed case in hyperbolic manifolds

In this section we show that even in the non-dynamically-coherent case, the mixed behavior is impossible for hyperbolic 3–manifolds. This will be done by using the study of translations in hyperbolic 3–manifolds developed in Sections 10 and 11 to provide more information on the dynamics of general partially hyperbolic diffeomorphisms.

The main result of this section is the following.

Theorem 12.1 *Let $f: M \to M$ be a partially hyperbolic diffeomorphism homotopic to the identity on a hyperbolic 3–manifold M. Suppose that there exists a finite lift and finite power $\hat{f}$ of f that preserves two branching foliations $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$, and is such that a good lift $\tilde{f}$ fixes a leaf of $\widetilde{\mathcal{W}}^{cu}$. Then f is a discretized Anosov flow.*

This, together with Proposition 6.1, completes Theorem 2.5.

12.1 The setup

Consider a partially hyperbolic diffeomorphism f as in Theorem 12.1.

Our goal is to show that the good lift $\tilde{f}$ of f fixes every leaf of $\widetilde{\mathcal{W}}^{cs}$ and $\widetilde{\mathcal{W}}^{cu}$. Indeed, Proposition 7.2 (and Corollary 7.5) then implies that $\hat{f}$ is dynamically coherent, so we can then use [3, Theorem B] to obtain that $\hat{f}$ is a discretized Anosov flow. In turn, thanks to Proposition 7.7, we obtain that f itself is dynamically coherent and a discretized Anosov flow.

Since Proposition 7.7 allows us to use finite lifts and powers, we assume directly that $f = \hat{f}$, that $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$ are orientable and transversely orientable, and that f preserves their orientations.

Since $\tilde{f}$ is assumed to fix one leaf of $\widetilde{\mathcal{W}}^{cu}$, Proposition 6.1 implies that every leaf of $\widetilde{\mathcal{W}}^{cu}$ is fixed. We will prove by contradiction that every leaf of $\widetilde{\mathcal{W}}^{cs}$ is fixed by $\tilde{f}$. So, by Proposition 6.1, we can assume that $\mathcal{W}^{cs}$ is $\mathbb{R}$–covered and uniform and that $\tilde{f}$ acts as a translation on the leaf space of $\widetilde{\mathcal{W}}^{cs}$. In particular, there are no center curves fixed by $\tilde{f}$.

Then, we can apply Proposition 5.2 to $\mathcal{W}^{cu}$ to deduce that every periodic center leaf is coarsely expanding.

On the other hand, since $\tilde{f}$ acts as a translation on $\widetilde{\mathcal{W}}^{cs}$, we can use the results from Sections 10 and 11. Let Φ^{cs} be a regulating pseudo-Anosov flow transverse to $\mathcal{W}^{cs}$ given by Proposition 10.1.

The flow Φ^{cs} is a genuine pseudo-Anosov, that is, it admits at least one periodic orbit which is a p–prong with $p \geq 3$; see [3, Proposition D.4].

Now, we choose γ in $\pi_1(M)$, associated to this prong, and apply Proposition 10.3: up to taking powers, we can assume that $h := \gamma \circ \widetilde{f}^k$ for some $k > 0$ fixes a leaf L of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Moreover, the dynamics in L resembles that of the dynamics of a p–prong, and in particular fixes every prong.

Notice that Proposition 11.1 also provides some center rays which are expanding in L for h. We will need to use some of the ideas involved in the proof of that proposition (even though the statement itself will not be used).

We summarize the discussion above in the following proposition.

Proposition 12.2 *Let $f\colon M \to M$ be a partially hyperbolic diffeomorphism of a hyperbolic 3–manifold M, homotopic to the identity, preserving branching foliations $\mathcal{W}^{\mathrm{cs}}$ and $\mathcal{W}^{\mathrm{cu}}$. Suppose that a good lift $\widetilde{f}$ fixes a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cu}}$ and acts as a translation on $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Then, up to taking finite iterates and covers, there exists $\gamma \in \pi_1(M)$ and $k > 0$ such that a center stable leaf $L \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$ is fixed by $h := \gamma \circ \widetilde{f}^k$, and its Lefschetz index is $I_{\mathrm{Fix}(h)}(h) = 1 - p$, with $p \geq 3$. Moreover, every center curve fixed by h in L is coarsely expanding.*

Let γ be as in the proposition. Let L be a center stable leaf fixed by $h = \gamma \circ \widetilde{f}^k$ and $L_2 = \widetilde{f}^k(L)$. As previously, we write $\tau_{12}\colon L \to L_2$ for the map obtained by flowing from L to L_2 along $\widetilde{\Phi}^{\mathrm{cs}}$. We set $g := \gamma \circ \tau_{12}$.

The map g acts on the compactification of L with its ideal circle $L \cup S^1(L)$ the same way as h does; see Sections 10 and 11.

Let δ be the unique orbit of $\widetilde{\Phi}^{\mathrm{cs}}$ fixed by γ and let x be the (unique) intersection of δ with L. Note that x is the unique fixed point of g. We assume that γ fixes the prongs of δ, so h has exactly $2p$ fixed points in $S^1(L)$. These fixed points are contracting if they correspond to an ideal point of $\mathcal{G}^{\mathrm{u}}(x)$, and expanding if they are ideal points of $\mathcal{G}^{\mathrm{s}}(x)$.

12.2 Proof of Theorem 12.1

To prove Theorem 12.1 we will first show some properties. Recall from Proposition 11.6 that every proper ray in $L \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$ fixed by h has a unique limit point in $S^1(L)$; notice that the ray must be either expanding or contracting. We will show that the fixed rays associated to the center and stable (branching) foliations have different limit points at infinity.

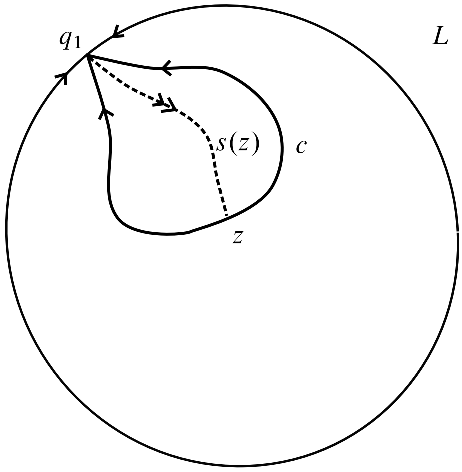

Figure 10: Rays have to land in different points of $S^1(L)$.

Lemma 12.3 *Let s be a stable leaf in L which is fixed by h. Then the two rays of s limit to distinct ideal points of L. The same holds if c is a center leaf in L fixed by h.*

Proof We do the proof for the center leaf c; the one for stable leaves is analogous, and a little bit easier (since there is no branching).

By hypothesis, c is fixed by h, hence it is coarsely expanding under h. It follows that there are fixed points of h in c. By Proposition 11.6, each ray of c can only limit to a point in $P \subset S^1(L)$ where, as previously, P is the set of attracting fixed points of h in $S^1(L)$. Let q_1 and q_2 be the ideal points of the rays. What we have to prove is that q_1 and q_2 are distinct.

Suppose that $q_1 = q_2$. Then c bounds a unique region S in L which limits only to $q_1 \in S^1(L)$. The other complementary region of c in L limits to every point in $S^1(L)$. Let z be a fixed point of h in c. Then the stable leaf $s(z)$ of z has a ray s_1 entering S. It cannot intersect c again, and it is properly embedded in L. Hence it has to limit to q_1 as well. See Figure 10.

But now this ray is contracting for h. This contradicts Proposition 11.6 because this ray should limit in a point of N. $\square$

Remark 12.4 The proof used strongly that periodic center leafs are coarsely expanding, in order to induce a behavior at infinity. In the examples of [7] it does happen that different stable curves land in the same ideal point at infinity in their center stable leaf.

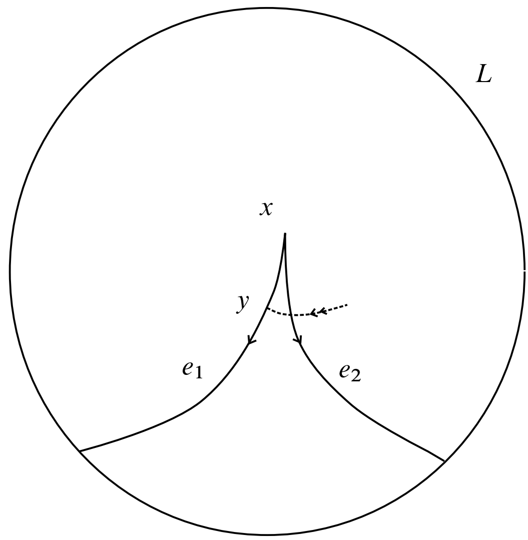

Figure 11: Showing the existence of fixed points below x in Lemma 12.5.

Now we show a sort of dynamical coherence for fixed center rays.

Lemma 12.5 *Suppose that c_1 and c_2 are distinct center leaves in L which are fixed by h. Then c_1 and c_2 cannot intersect.*

Notice that since f is not necessarily dynamically coherent, the distinct center leaves c_1 and c_2 can a priori intersect each other. The proof will depend very strongly on the fact that center rays fixed by h are coarsely expanding.

Proof Suppose that c_1 and c_2 intersect. Since c_1 and c_2 are both fixed by h, so is their intersection. Since h is coarsely expanding in each, c_1 and c_2 share a fixed point of h. In the proof of Claim 11.3, we showed that c_1 and c_2 cannot form a bigon B.

It follows that there is a point x, fixed by h, which is an endpoint of all intersections of c_1 and c_2: on one side x bounds a ray e_1 of c_1 and a ray e_2 of c_2 such that e_1 and e_2 are disjoint. For a point y in e_1 near enough to x, we have that $s(y)$ must intersect c_2. Since stable lengths are contracting under powers of h, it implies that e_1 is contracting towards x near x and similarly for e_2; see Figure 11. But e_1 is coarsely expanding. Hence there must exist fixed points of h in e_1. Let $y \in e_1$ be the closest point to x which is fixed by h. Similarly, let z in e_2 be the closest to x fixed by h.

The leaves $s(y)$ and $s(z)$ are not separated from each other in the stable leaf space in L.

Let now c be a center leaf through x which is between c_1 and c_2 and which is the first center leaf not intersecting $s(y)$.

Then $h(c) = c$ since $s(y)$ is fixed and c is the first leaf through x not intersecting $s(y)$. Consider the ray of c starting at x and moving in the direction of y. This ray is the limit of compact center segments from x to points in $s(y)$. As such this ray of c can only intersect stable leaves which are between $s(x)$ and $s(y)$. Because the map h contracts stable lengths it follows that the map h is contracting in this ray of c. This contradicts Proposition 12.2 because this ray is in a center leaf which is fixed by h. □

Thus far, we showed that distinct center leaves in L which are fixed by h do not intersect. Then the proof of Claim 11.4 also implies that fixed center leaves cannot accumulate (as accumulation would imply that some fixed leaves intersect).

We conclude that there are finitely many center leaves in L that are fixed under h. Each such center leaf is coarsely expanding. For each such center leaf c, we consider a small enough open topological disk containing all the fixed points of h in c, and no other fixed point of h in L. Then, on such disks, the Lefschetz index of h is -1. Since the total Lefschetz number of h in L is $1 - p$ it follows that:

Lemma 12.6 *There are exactly $p - 1$ center leaves which are fixed by h in L.*

This together with the following lemma will allow us to make a counting argument to reach a contradiction.

Lemma 12.7 *Let c_1 and c_2 be two distinct center leaves in L fixed by h. Let $y_1 \in c_1$ and $y_2 \in c_2$ be fixed points of h. Then $s(y_1)$ and $s(y_2)$ do not have common ideal points.*

Proof Suppose, for a contradiction, that there are distinct fixed center leaves c_1 and c_2 satisfying the following: there are points $y_1 \in c_1$ and $y_2 \in c_2$, fixed by h, such that $s_1 = s(y_1)$ and $s_2 = s(y_2)$ share an ideal point in $S^1(L)$.

Let q be the common ideal point of the corresponding rays of s_1 and s_2. Note that by Proposition 12.2 the point q cannot be an endpoint of c_1 or c_2, because ideal points of fixed centers are contracting in $S^1(L)$ and ideal points of fixed stables are repelling in $S^1(L)$.

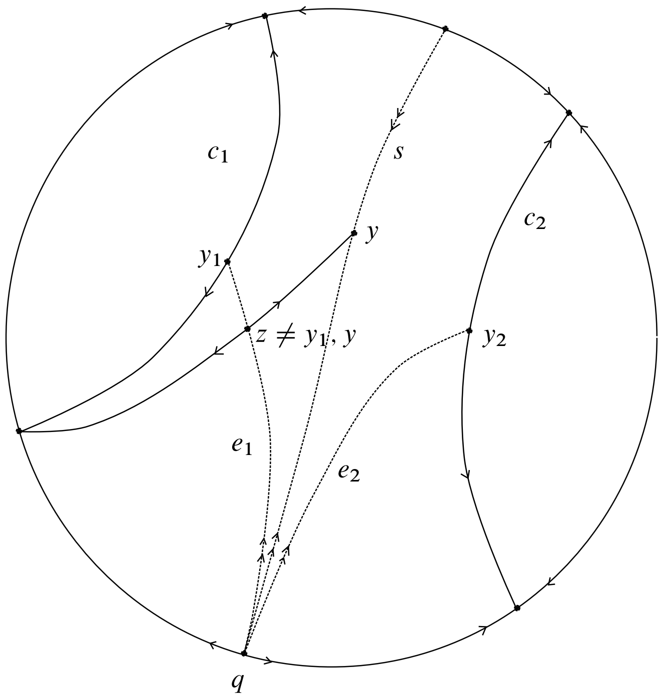

Figure 12: A depiction of the main objects in the proof of Lemma 12.7.

Let e_j be the ray in s_j with endpoint y_j and ideal point q. Suppose first that no center leaf intersecting e_1 intersects e_2. Let c_0 be a center leaf intersecting e_1. Iterate c_0 by powers of h^{-1}. It pushes points in s_1 away from y_1. Since the leaves $h^{-i}(c_0)$ all intersect s_1 and none of them intersects s_2 or c_2, the sequence $(h^{-i}(c_0))$ converges to a collection of center leaves as $i \to +\infty$. Then there is only one center leaf in this limit, call it c, which separates all of $h^{-i}(c_0)$ from s_2. This c is invariant under h and it has an ideal point in q because it separates $h^{-i}(c_0)$ (recall that $h^{-i}(c_0) \cap s_1 \to q$ as $i \to \infty$) from s_2. Now q is a repelling fixed point in $S^1(L)$, so c must have an attracting ray, a contradiction with Proposition 12.2.

It follows that some center leaf intersecting e_1 also intersects e_2. Let c_0 be one such center leaf. Now iterate by positive powers of h. Then $(h^i(c_0))$ converges to a fixed center leaf v_1 through y_1 and a fixed center leaf v_2 through y_2. But then v_1 and c_1 are both fixed by h and both contain y_1. Lemma 12.5 implies that $c_1 = v_1$ and $c_2 = v_2$. In particular $v_1 \neq v_2$, and they are nonseparated from each other. In this case, consider s the unique stable leaf defined as the first leaf not intersecting c_1 that separates s_1 from s_2. Then, as above, h fixes s and has a fixed point y in s. But a center leaf c

through y fixed by h has to intersect the interior of the ray e_1. This intersection point is the intersection of c fixed by h, and s_1 fixed by h. So this intersection point is fixed by h. But this is a contradiction, because y_1 is the only fixed point of h in s_1. So Lemma 12.7 is proven. □

We now can complete the proof of Theorem 12.1.

Proof of Theorem 12.1 By Lemma 12.6, there are $p-1$ center leaves fixed by h in L. We denote them by $c_1, \dots, c_{p-1}$.

Each center leaf has at least one fixed point. Let y_i, for $1 \le i \le p-1$, be a fixed point in c_i. Then, for each i, Lemma 12.3 states that $s(y_i)$ has two distinct ideal points z_i^1 and z_i^2.

Moreover, for every $i \neq j$, the ideal points of the stable leaves are distinct by Lemma 12.7. It follows that there are at least $2p-2$ distinct points in $S^1(L)$ which are repelling.

But we also know that there are exactly p points in $S^1(L)$ that are repelling under h. It follows that $2p-2 \le p$, which implies $p=2$. However, we had that $p \ge 3$, thus obtaining a contradiction.

This finishes the proof of Theorem 12.1. □

Appendix A Some 3–manifold topology

Besides the 3–manifold topology presented in [3, Appendix A] we will need an additional result, which is important for understanding certain particular deck transformations when one lifts to finite covers.

Lemma A.1 *Let M be a closed, irreducible 3–manifold with fundamental group that is not virtually nilpotent. Suppose that β is a nontrivial deck transformation so that $d(x, \beta(x))$ is bounded above in $\widetilde{M}$. Then M is a Seifert fibered space and β represents a power of a regular fiber.*

Proof First we assume that M is orientable. Then, the JSJ decomposition states that M has a canonical decomposition into Seifert fibered and geometrically atoroidal pieces. We lift this to a decomposition of $\widetilde{M}$ and construct a tree $\mathcal{T}$ in the following way: the vertices are the lifts of components of the torus decomposition of M, and we associate an edge if two components intersect along the lift of a torus. Such a lift of a torus is called a wall. There is a minimum separation distance between any two walls.

The deck transformation β acts on this tree. Let W be a wall. Suppose that $\beta(W)$ is distinct from W. But, as subsets of $\widetilde{M}$, the walls W and $\beta(W)$ are a finite Hausdorff distance from each other. Then $\pi(W)$ and $\pi(\beta(W))$ are tori in M, and the region V in $\widetilde{M}$ between W and $\beta(W)$ projects to $\pi(V)$, which is $\mathbb{T}^2\times[0,1]$ in M. If this happens, then M is a torus bundle over a circle. In that case, use that $\pi_1(M)$ is not virtually nilpotent, so the monodromy of the fibration is an Anosov map of $\mathbb{T}^2$. But then no β as above could satisfy the bounded distance property. It follows that $\beta(W) = W$ for any wall, and in particular $\beta(P) = P$ for any vertex of $\mathcal{T}$.

Now consider a vertex P. Suppose first that $\pi(P)$ is homotopically atoroidal. By the geometrization theorem, $\pi(P)$ is hyperbolic. If β restricted to P were to satisfy the bounded distance property, then it would have to be the identity on P. Hence β itself is the identity, a contradiction.

Hence all the pieces of the torus decomposition of M are homotopically toroidal. Suppose now that there is one such piece $\pi(P)$ that is geometrically atoroidal (but not homotopically atoroidal). The proof of the Seifert fibered conjecture [16; 20] shows that $\pi(P)$ has no boundary and $\pi(P)$ is Seifert. In other words, $M = \pi(P)$ is Seifert. So we can assume that all the pieces of the torus decomposition are geometrically toroidal. Then they are all Seifert fibered. Thus M is a graph manifold.

We will show that the torus decomposition of M is in fact trivial, proving that M is Seifert fibered. Suppose it is not true. Then the tree $\mathcal{T}$ is infinite. Let P_1, P_2, P_3 be three consecutive vertices in $\mathcal{T}$. Let W_1 be the wall between P_1 and P_2. Then $\beta(W_1)$ (as a set in $\widetilde{M}$) is a bounded distance from W_1 and sends the Seifert fibration of P in W_1 to lifts of Seifert fibers. It follows that $\beta = \delta_1^k\alpha_1$, where δ_1 represents a regular fiber in $\pi(P_1)$, and α_1 is a loop in $\pi(W_1)$. Similarly, if W_2 is the wall between P_2 and P_3, then $\beta = \delta_3^i\alpha_3$, where α_3 is a loop in $\pi(W_3)$. Then α_1 and α_3 are both in the boundary of $\pi(P_2)$. The loops representing $\delta_1^k\alpha_1$ and $\delta_3^i\alpha_3$ are both in the boundary of $\pi(P_2)$. They represent the same element of $\pi_1(M)$ only when $k = i = 0$ and α_1 and α_3 are freely homotopic. That means that P_2 is a torus times an interval, which is impossible in the torus decomposition in our situation, as explained above.

It follows now that the torus decomposition of M is trivial, which implies that M is Seifert fibered. Moreover, if the base is not hyperbolic, then $\pi_1(M)$ is virtually nilpotent [32, Theorem 5.3]. But this contradicts the hypothesis of the lemma.

It follows that the base is hyperbolic. Also β induces a transformation in the universal cover of the base that is a bounded distance from the identity. This can only happen if this transformation is the identity. Therefore β represents a power of a regular Seifert fiber in M (notice that nonregular fibers induce a finite symmetry on the base, thus not the identity, and not a bounded distance from the identity).

So the lemma is proven when M is orientable. If M is not orientable, then it has a double cover M_2 which is orientable. Now β^2 lifts to an element of $\pi_1(M_2)$ that satisfies the assumption of the lemma. So we can apply the result to M_2 and obtain that M_2 is Seifert. Thus M is doubly covered by a Seifert space, which, by a result of Tollefson [34], implies that M itself is Seifert fibered. It follows that β corresponds to a power of a regular fiber. This finishes the proof of the lemma. □

Appendix B Minimality and f–minimality

We prove that in certain situations minimality is equivalent to f–minimality. We need the following result, which is of interest in itself.

Lemma B.1 *Let $\mathcal{L}^{\mathrm{cs}}$ be the leaf space of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. Let $\mathcal{B} \subset \mathcal{L}^{\mathrm{cs}}$ be a closed set of leaves. Suppose that, for all $x \in \widetilde{M}$, there exists a leaf $L \in \mathcal{B}$ containing x. Then $\mathcal{B} = \mathcal{L}^{\mathrm{cs}}$.*

Proof The lemma is obvious when $\mathcal{W}^{\mathrm{cs}}$ is a true foliation (and one does not need to require $\mathcal{B}$ to be closed). However, when $\mathcal{W}^{\mathrm{cs}}$ has some branching, one could possibly have a union of leaves that cover all of $\widetilde{M}$ without using all the leaves of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$. For closed sets of leaves we show this is not possible.

Let L be a leaf of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, x a point in L and τ an open unstable segment through x. The set of leaves of $\widetilde{\mathcal{W}}^{\mathrm{cs}}$ intersecting τ is isomorphic to an open interval. Using the transversal orientation to $\widetilde{\mathcal{W}}^{\mathrm{cs}}$, we can put an order on this interval.

By our assumption, every point in τ intersects a leaf in $\mathcal{B}$. Let L' be the supremum of leaves in $\mathcal{B}$, intersecting τ and smaller than or equal to L. Since $\mathcal{B}$ is closed, we have $L' \in \mathcal{B}$. Notice that x is in both L and L'.

We claim that $L' = L$. If L is not equal to L' then they branch out. Let y be a boundary point of $L \cap L'$. Let $z \in L'$, with $z \notin L$ close enough to y that its unstable leaf $u(z)$ intersects L. Now take any point $w \in u(z)$ in between z and $L \cap u(z)$. Any leaf $L_1 \in \widetilde{\mathcal{W}}^{\mathrm{cs}}$ that contains w must contain y. Hence (because leaves do not cross), L_1 also contains x. By definition, it is above L', thus L_1 is not in $\mathcal{B}$. Since this is true for any leaf through w, it contradicts our assumption. □

Lemma B.2 *When $\mathcal{W}^{cs}$ does not have compact leaves, then f–minimality of $\mathcal{W}^{cs}$ is equivalent to minimality of $\mathcal{W}^{cs}$.*

Proof Minimality obviously implies f–minimality, so we only need to show the other implication.

Suppose that $\mathcal{W}^{cs}$ is not minimal and let C be the union of a set of $\mathcal{W}^{cs}$ leaves which is closed and not M. Let $\mathcal{W}^{cs}_\epsilon$ be an approximating foliation, with approximating map h^{cs}_ϵ sending leaves of $\mathcal{W}^{cs}_\epsilon$ to those of $\mathcal{W}^{cs}$. Then $(h^{cs}_\epsilon)^{-1}(C)$ is a set which is a union of $\mathcal{W}^{cs}_\epsilon$ leaves, which is closed and not M. In particular, it contains an exceptional minimal set D. By [26, Theorem 4.1.3], the actual foliation $\mathcal{W}^{cs}_\epsilon$ has finitely many exceptional minimal sets $B_1, \dots, B_k$. The union B of these is not M because $D \neq M$. The set of leaves in B is a closed set of leaves denoted by $\mathcal{B}$. Then $A = h^{cs}_\epsilon(B)$ is a closed subset of M, and $\mathcal{A} = h^{cs}_\epsilon(\mathcal{B})$ is a closed set of leaves, being the image by h^{cs}_ϵ of the leaves in $\mathcal{B}$. Let $\widetilde{\mathcal{A}} = \pi^{-1}(\mathcal{A})$; we stress that this is on the leaf-space level, not in terms of sets. This is a closed subset of $\mathcal{L}^{cs}$.

Let $A_i := h^{cs}_\epsilon(B_i)$. Every leaf of $\mathcal{W}^{cs}$ which is the image of a leaf in B_i is dense in A_i. Using this, it is easy to see that $f(A) = A$. By f–minimality it follows that $A = M$.

Since $A = M$, $\widetilde{\mathcal{A}}$ is a closed subset of $\mathcal{L}^{cs}$, whose union of points in all leaves of $\widetilde{\mathcal{A}}$ is $\widetilde{M}$, as $A = M$. Lemma B.1 implies that $\widetilde{\mathcal{A}} = \mathcal{L}^{cs}$. Hence for each leaf E of $\mathcal{W}^{cs}$, it is the image of a leaf F in some B_i. Conversely, every leaf of $\mathcal{W}^{cs}_\epsilon$ maps by h^{cs}_ϵ to a leaf of $\mathcal{W}^{cs}$.

For each leaf E of $\mathcal{W}^{cs}$, its preimage $(h^{cs}_\epsilon)^{-1}(E)$ is a closed interval of leaves of $\mathcal{W}^{cs}_\epsilon$. No leaf in the interior of the interval can be in a B_i as it is a minimal set. It follows that the complementary regions to B in M are I–bundles. These can be collapsed to generate another foliation $\mathcal{C}$. Since the B_i were minimal sets of $\mathcal{W}^{cs}_\epsilon$, the collapsing of each of these is a minimal set of $\mathcal{C}$. Since the union is all of M, there can be only one such minimal set, so $\mathcal{W}^{cs}_\epsilon$ is minimal.

But this contradicts the fact that D is an exceptional minimal set of $\mathcal{W}^{cs}_\epsilon$. $\square$

We state the following criteria for dynamical coherence (which in this setting is quite obvious).

Proposition B.3 [8, Proposition 1.6 and Remark 1.10] *Assume that f is a partially hyperbolic diffeomorphism admitting branching foliations $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$. If no two distinct leaves of $\mathcal{W}^{cs}$ or $\mathcal{W}^{cu}$ intersect, then f is dynamically coherent.*

Appendix C The Lefschetz index

Here we define the Lefschetz index and give the main property that we used. We refer to the monograph by Franks [19, Section 5] for details and other references.

For any space X and subset $A \subset X$, we denote by $H_k(X, A)$ the k^{th} relative homology group with coefficients in $\mathbb{Z}$.

Definition C.1 Let $V \subset \mathbb{R}^k$ be an open set and $F\colon V \subset \mathbb{R}^k \to \mathbb{R}^k$ be a continuous map such that the set of fixed points of F is $\Gamma \subset V$, a compact set. Then the *Lefschetz index* of F, denoted by $I_\Gamma(F)$, is an element in $\mathbb{Z} \cong H_k(\mathbb{R}^k, \mathbb{R}^k - \{0\})$, defined as follows. It is the image by $(\mathrm{id} - F)_*\colon H_k(V, V - \Gamma) \to H_k(\mathbb{R}^k, \mathbb{R}^k - \{0\})$ of the class u_Γ, where u_Γ itself is the image of the generator 1 under the composite $H_k(\mathbb{R}^k, \mathbb{R}^k - D) \to H_k(\mathbb{R}^k, \mathbb{R}^k - \Gamma) \cong H_k(V, V - \Gamma)$. Here D is a ball containing Γ.

It is easy to see that if $\Gamma = \mathrm{Fix}(F) = \Gamma_1 \cup \cdots \cup \Gamma_j$, where Γ_i are compact and disjoint, then $I_\Gamma(F) = \sum_1^j I_\Gamma(F)$. Here $I_\Gamma(F)$ is the index restricted to an open set V_i of V which does not intersect the other Γ_m; see [19, Theorem 5.8(b)].

This technical definition works well with the standard examples. For a single hyperbolic fixed point q, the index at q is exactly $\mathrm{sgn}(\det(\mathrm{id} - D_q F))$, where det is the determinant and sgn is the sign of the determinant; see [19, Proposition 5.7]. Hence in dimension two, the index of a hyperbolic fixed point when the orientation of the bundles is preserved is -1. This can be generalized to a p–prong hyperbolic fixed point, to obtain that the index is $1 - p$. This is because the index is invariant under homotopic changes. A p–prong can be easily split into $p - 1$ distinct hyperbolic points which are differentiable. In addition, for any fixed set which behaves locally as a hyperbolic fixed point, the index is the same as the hyperbolic fixed point.

The main property we use is the following.

Proposition C.2 [19, Theorem 5.8(c)] *Let P be a topological plane equipped with a metric d. Let $g, h\colon P \to P$ be two homeomorphisms. Suppose that there exists $R > 0$ such that*

- *for every $x \in P$, one has that $d(g(x), h(x)) < R$, and*
- *there is a disk D such that, for every $x \notin D$, one has that $d(x, g(x)) > 2R$.*

Then the total index satisfies $I_{\mathrm{Fix}(g)}(g) = I_{\mathrm{Fix}(h)}(h)$.

See also [28, Section 8.6] for an alternative presentation of the Lefschetz index.

References

[1] **T Barbot**, *Actions de groupes sur les 1–variétés non séparées et feuilletages de codimension un*, Ann. Fac. Sci. Toulouse Math. 7 (1998) 559–597 MR Zbl

[2] **M Barge**, **J Franks**, *Recurrent sets for planar homeomorphisms*, from "From topology to computation: proceedings of the Smalefest" (M W Hirsch, J E Marsden, M Shub, editors), Springer (1993) 186–195 MR Zbl

[3] **T Barthelmé**, **S R Fenley**, **S Frankel**, **R Potrie**, *Partially hyperbolic diffeomorphisms homotopic to the identity in dimension 3, I: The dynamically coherent case*, preprint (2019) arXiv To appear in Ann. Sci. École Norm. Sup.

[4] **T Barthelmé**, **S R Fenley**, **S Frankel**, **R Potrie**, *Dynamical incoherence for a large class of partially hyperbolic diffeomorphisms*, Ergodic Theory Dynam. Systems 41 (2021) 3227–3243 MR Zbl

[5] **M Boileau**, **J-P Otal**, *Scindements de Heegaard et groupe des homéotopies des petites variétés de Seifert*, Invent. Math. 106 (1991) 85–107 MR Zbl

[6] **C Bonatti**, **L J Díaz**, **M Viana**, *Dynamics beyond uniform hyperbolicity: a global geometric and probabilistic perspective*, Encyclopaedia of Mathematical Sciences 102, Springer (2005) MR Zbl

[7] **C Bonatti**, **A Gogolev**, **A Hammerlindl**, **R Potrie**, *Anomalous partially hyperbolic diffeomorphisms, III: Abundance and incoherence*, Geom. Topol. 24 (2020) 1751–1790 MR Zbl

[8] **C Bonatti**, **A Wilkinson**, *Transitive partially hyperbolic diffeomorphisms on 3–manifolds*, Topology 44 (2005) 475–508 MR Zbl

[9] **M Brin**, **D Burago**, **S Ivanov**, *Dynamical coherence of partially hyperbolic diffeomorphisms of the 3–torus*, J. Mod. Dyn. 3 (2009) 1–11 MR Zbl

[10] **D Burago**, **S Ivanov**, *Partially hyperbolic diffeomorphisms of 3–manifolds with abelian fundamental groups*, J. Mod. Dyn. 2 (2008) 541–580 MR Zbl

[11] **D Calegari**, *The geometry of $\mathbb{R}$–covered foliations*, Geom. Topol. 4 (2000) 457–515 MR Zbl

[12] **D Calegari**, *Foliations and the geometry of 3–manifolds*, Oxford Univ. Press (2007) MR Zbl

[13] **A Candel**, *Uniformization of surface laminations*, Ann. Sci. École Norm. Sup. 26 (1993) 489–516 MR Zbl

[14] **A Candel**, **L Conlon**, *Foliations, I*, Graduate Studies in Mathematics 23, Amer. Math. Soc., Providence, RI (2000) MR Zbl

[15] **P D Carrasco**, **F Rodriguez-Hertz**, **J Rodriguez-Hertz**, **R Ures**, *Partially hyperbolic dynamics in dimension three*, Ergodic Theory Dynam. Systems 38 (2018) 2801–2837 MR Zbl

[16] **A Casson**, **D Jungreis**, *Convergence groups and Seifert fibered* 3*–manifolds*, Invent. Math. 118 (1994) 441–456 MR Zbl

[17] **S R Fenley**, *Foliations, topology and geometry of* 3*–manifolds:* $\mathbb{R}$*–covered foliations and transverse pseudo-Anosov flows*, Comment. Math. Helv. 77 (2002) 415–490 MR Zbl

[18] **S R Fenley**, **R Potrie**, *Ergodicity of partially hyperbolic diffeomorphisms in hyperbolic* 3*–manifolds*, Adv. Math. 401 (2022) art. id. 108315 MR Zbl

[19] **J M Franks**, *Homology and dynamical systems*, CBMS Regional Conference Series in Mathematics 49, Amer. Math. Soc., Providence, RI (1982) MR Zbl

[20] **D Gabai**, *Convergence groups are Fuchsian groups*, Ann. of Math. 136 (1992) 447–510 MR Zbl

[21] **D Gabai**, **G R Meyerhoff**, **N Thurston**, *Homotopy hyperbolic* 3*–manifolds are hyperbolic*, Ann. of Math. 157 (2003) 335–431 MR Zbl

[22] **A Hammerlindl**, **R Potrie**, *Pointwise partial hyperbolicity in three-dimensional nilmanifolds*, J. Lond. Math. Soc. 89 (2014) 853–875 MR Zbl

[23] **A Hammerlindl**, **R Potrie**, *Classification of partially hyperbolic diffeomorphisms in* 3*–manifolds with solvable fundamental group*, J. Topol. 8 (2015) 842–870 MR Zbl

[24] **A Hammerlindl**, **R Potrie**, *Partial hyperbolicity and classification: a survey*, Ergodic Theory Dynam. Systems 38 (2018) 401–443 MR Zbl

[25] **A Hammerlindl**, **R Potrie**, **M Shannon**, *Seifert manifolds admitting partially hyperbolic diffeomorphisms*, J. Mod. Dyn. 12 (2018) 193–222 MR Zbl

[26] **G Hector**, **U Hirsch**, *Introduction to the geometry of foliations, B: Foliations of codimension one*, 2nd edition, Aspects of Mathematics E3, Vieweg & Sohn, Braunschweig (1987) MR Zbl

[27] **F R Hertz**, **J R Hertz**, **R Ures**, *Center-unstable foliations do not have compact leaves*, Math. Res. Lett. 23 (2016) 1819–1832 MR Zbl

[28] **A Katok**, **B Hasselblatt**, *Introduction to the modern theory of dynamical systems*, Encyclopedia of Mathematics and its Applications 54, Cambridge Univ. Press (1995) MR Zbl

[29] **R Potrie**, *Partial hyperbolicity and foliations in* $\mathbb{T}^3$, J. Mod. Dyn. 9 (2015) 81–121 MR Zbl

[30] **R Potrie**, *Robust dynamics, invariant structures and topological classification*, from "Proceedings of the International Congress of Mathematicians, III: Invited lectures" (B Sirakov, P N de Souza, M Viana, editors), World Sci., Rio de Janeiro, 2018 (2018) 2063–2085 MR Zbl

[31] **F Rodriguez Hertz**, **M A Rodriguez Hertz**, **R Ures**, *A non-dynamically coherent example on* $\mathbb{T}^3$, Ann. Inst. H. Poincaré C Anal. Non Linéaire 33 (2016) 1023–1032 MR Zbl

[32] **P Scott**, *The geometries of 3–manifolds*, Bull. London Math. Soc. 15 (1983) 401–487 MR Zbl

[33] **W P Thurston**, *Three-manifolds, foliations and circles, I*, preprint (1997) arXiv

[34] **J L Tollefson**, *Involutions of Seifert fiber spaces*, Pacific J. Math. 74 (1978) 519–529 MR Zbl

[35] **F Waldhausen**, *On irreducible 3–manifolds which are sufficiently large*, Ann. of Math. 87 (1968) 56–88 MR Zbl

Department of Mathematics and Statistics, Queen's University
Kingston, ON, Canada

Department of Mathematics, Florida State University
Tallahassee, FL, United States

Department of Mathematics, Washington University in St. Louis
St Louis, MO, United States

Facultad de Ciencias – Centro de Matemática, Universidad de la República
Montevideo, Uruguay

thomas.barthelme@queensu.ca, fenley@math.fsu.edu,
steven.frankel@wustl.edu, rpotrie@cmat.edu.uy

sites.google.com/site/thomasbarthelme,
http://www.cmat.edu.uy/~rpotrie/

Proposed: David Gabai Received: 7 August 2020
Seconded: Leonid Polterovich, David Fisher Revised: 4 February 2022

msp

Geometry & Topology **27**:8 (2023) 3183–3228
DOI: 10.2140/gt.2023.27.3183
Published: 9 November 2023

The Weil–Petersson gradient flow of renormalized volume and 3–dimensional convex cores

MARTIN BRIDGEMAN
JEFFREY BROCK
KENNETH BROMBERG

We use the Weil–Petersson gradient flow for renormalized volume to study the space $CC(N; S, X)$ of convex cocompact hyperbolic structures on the relatively acylindrical 3–manifold $(N; S)$. Among the cases of interest are the deformation space of an acylindrical manifold and the Bers slice of quasifuchsian space associated to a fixed surface. To treat the possibility of degeneration along flow-lines to peripherally cusped structures, we introduce a surgery procedure to yield a surgered gradient flow that limits to the unique structure $M_{\text{geod}} \in CC(N; S, X)$ with totally geodesic convex core boundary facing S. Analyzing the geometry of structures along a flow line, we show that if $V_{\boldsymbol{R}}(M)$ is the renormalized volume of M, then $V_{\boldsymbol{R}}(M) - V_{\boldsymbol{R}}(M_{\text{geod}})$ is bounded below by a linear function of the Weil–Petersson distance $d_{\text{WP}}(\partial_c M, \partial_c M_{\text{geod}})$, with constants depending only on the topology of S. The surgered flow gives a unified approach to a number of problems in the study of hyperbolic 3–manifolds, providing new proofs and generalizations of well-known theorems such as Storm's result that M_{geod} has minimal volume for N acylindrical and the second author's result comparing convex core volume and Weil–Petersson distance for quasifuchsian manifolds.

32G15, 30F40, 30F60; 32Q45, 51P05

1 Introduction

The use of a *geometric flow*, or a flow on a space of metrics on a given manifold, has provided an abundantly fruitful approach to understanding a manifold's structure. In our previous work [4], we introduced a new geometric flow on the space of *hyperbolic* metrics on a 3–manifold that admits a hyperbolic structure, showing how the flow can be used to discover the metric of least *convex core volume*. In the present paper, we illustrate how this flow provides an analytic version of results on convex core volume that were available previously only through combinatorial methods, demonstrating how this approach allows for conjectured extensions to much more general cases.

When a hyperbolic 3–manifold M admits a compact convex submanifold we say it is *convex cocompact*; the geometry of the smallest such submanifold, its *convex core*, carries all the interesting information about its geometry. For such M (or more generally conformally compact Einstein manifolds), work of Graham and Witten [17] in physics led to an alternative notion of *renormalized volume*. From a mathematical perspective, this concept has been elaborated in a series of papers of Krasnov and Schlenker [21; 22], Takhtajan and Teo [32] and Zograf and Takhtajan [35]. The renormalized volume $V_R(M)$ of M connects many analytic notions from the deformation theory to the geometry of M and is closely related to classical objects such as the convex core volume $V_C(M)$ and the Weil–Petersson geometry of Teichmüller space.

If N is a compact 3–manifold admitting a complete hyperbolic structure of finite volume, the renormalized volume gives an analytic function $V_R\colon CC(N) \to \mathbb{R}$, where $CC(N)$ is the deformation space of convex cocompact structures on N. We will give a precise definition of V_R later in the paper, but knowledge of its basic properties will be largely sufficient for our purposes. In particular, the differential dV_R on $CC(N)$ is described in terms of the classical *Schwarzian derivative* and can be used as a definition of V_R.

A convex cocompact structure $M \in CC(N)$ is naturally compactified by a *complex projective structure* on ∂N. The underlying conformal structure is the *conformal boundary* $\partial_c M$ of M. The Schwarzian derivative associated to the projective structure determines a holomorphic quadratic differential $\phi_M \in Q(\partial_c M)$. The utility of the renormalized volume function lies in a particularly clean formula for its derivative, first shown by Takhtajan and Zograf [35] and Takhtajan and Teo [32]. A new proof was given by Krasnov and Schlenker [22, Lemma 8.5] using methods that are more closely aligned with the present work. To state the result, we recall that $CC(N)$ is (locally) parametrized by $\mathrm{Teich}(\partial N)$, and the cotangent space at $\partial_c M$ is parametrized by $Q(\partial_c M)$. We then have:

Theorem 1.1 [35; 32; 22] *Let μ be an infinitesimal Beltrami differential on $\partial_c M$. Then*

$$dV_R(\mu) = \mathrm{Re} \int_{\partial_c M} \phi_M \mu.$$

By integrating this formula along a Weil–Petersson geodesic and applying the classical Kraus–Nehari bound on the L^∞–norm of ϕ_M, Schlenker [29, Theorem 1.2] obtained the following for the quasifuchsian structure $Q(X,Y)$ on $N = S\times[0,1]$ with conformal boundary $X \sqcup Y$:

$$V_R(Q(X,Y)) \le 3\sqrt{\frac{\pi}{2}|\chi(S)|}\, d_{\mathrm{WP}}(X,Y).$$

Furthermore, Schlenker showed that for quasifuchsian manifolds, the renormalized volume and the volume of the convex core are boundedly related. A more refined version (see [4, Theorems 2.16 and 3.7]) is

$$V_C(Q(X,Y)) - 6\pi|\chi(S)| \le V_R(Q(X,Y)) \le V_C(Q(X,Y)).$$

Combined, these gave a new proof of an upper bound on the volume the convex core of $Q(X,Y)$ in terms of $d_{\mathrm{WP}}(X,Y)$ originally due to the second author [9], resulting also in new approaches to the study of volumes of fibered 3–manifolds in [11; 19] generalizing and sharpening known estimates [10].

Here, the variational formula (Theorem 1.1) will be our jumping-off point to study the Weil–Petersson gradient flow of V_R. It will be useful to restrict V_R to certain subspaces of the space of convex cocompact structures $CC(N)$. In particular, let $(N;S)$ be a pair where N is a compact hyperbolizable 3–manifold and $S \subseteq \partial N$ is a collection of components of the boundary. Then $CC(N;S,X) \subseteq CC(N)$ is the space of convex cocompact hyperbolic structures on N where the conformal boundary on the complement of S is the fixed conformal structure X. The pair $(N;S)$ is *relatively incompressible* if the inclusion $S \hookrightarrow N$ is π_1–injective, and *relatively acylindrical* if there are no essential cylinders with boundary in S. Note that the second condition implies the first.

In this paper our focus will be on when $(N;S)$ is relatively acylindrical. The cases of greatest interest are

(1) when $S = \partial N$, and N itself is acylindrical, and

(2) when $N = S\times[0,1]$, and $CC(N; S\times\{1\}, X)$ is a *Bers slice* of the space of quasifuchsian structures.

One important feature of relatively acylindrical pairs is that the deformation space $CC(N;S,X)$ has a unique hyperbolic structure M_{geod} where the components of the convex core facing S are totally geodesic. The main application of our study of the gradient flow is the following.

Theorem A *Let $CC(N,S;X)$ be a relatively acylindrical deformation space. There exists $A(S)$, depending only on the topology of S, and a universal constant δ such that*

$$A(S)\big(d_{\mathrm{WP}}(\partial_c M_{\mathrm{geod}}, \partial_c M) - \delta\big) \le V_{\boldsymbol{R}}(M) - V_{\boldsymbol{R}}(M_{\mathrm{geod}}).$$

For a Bers slice $CC(S\times[0,1], S\times\{1\}, X)$, we have $M_{\mathrm{geod}} = Q(X,X)$ and both the convex core and renormalized volume of this Fuchsian manifold are zero. Applying the above comparison between renormalized volume and convex core volume, we obtain:

Theorem B *Let S be a closed surface of genus $g \ge 2$. Then we have*

$$A(S)\big(d_{\mathrm{WP}}(X,Y) - \delta\big) \le V_C(Q(X,Y)) \le 3\sqrt{\frac{\pi}{2}|\chi(S)|}\, d_{\mathrm{WP}}(X,Y) + 6\pi|\chi(S)|.$$

Schlenker's argument in the quasifuchsian case also applies to relatively acylindrical manifolds, so we have for any M and M' in $CC(N;S,X)$ that

$$V_{\boldsymbol{R}}(M) - V_{\boldsymbol{R}}(M') \le 3\sqrt{\frac{\pi}{2}|\chi(S)|}\, d_{\mathrm{WP}}(\partial_c M, \partial_c M').$$

If we let $M_{\mathrm{geod}} = M'$, then we get an upper bound on the expression in Theorem A. The comparison between renormalized volume and convex core volume also extends to acylindrical manifolds (or any manifold with incompressible boundary).

Theorem C *Let N be a hyperbolizable, acylindrical 3–manifold. Then*

$$\begin{aligned}A(\partial N)\big(d_{\mathrm{WP}}(\partial_c M_{\mathrm{geod}}, \partial_c M) - \delta\big) &\le V_C(M) - V_C(M_{\mathrm{geod}}) \\ &\le 3\sqrt{\frac{\pi}{2}|\chi(\partial N)|}\, d_{\mathrm{WP}}(\partial_c M_{\mathrm{geod}}, \partial_c M) + 3\pi|\chi(\partial N)|,\end{aligned}$$

where A and δ are as in Theorem A.

Remark The constants in Theorem C *depend only on the topology of* ∂N. While we expect the second author's original method combined with Thurston's compactness theorem for hyperbolic structures on acylindrical manifolds should also produce a similar bound, the constants in such an approach would depend on the topology of N, due to the application of Thurston's result. The approach taken here is thus not only more direct but produces a stronger result. In particular, while Thurston's compactness theorem implies that the convex core of M_{geod} has a bi-Lipschitz embedding into any complete hyperbolic structure on N where the bi-Lipschitz constants only depend on N, it is natural to conjecture that these bi-Lipschitz constants only depend on ∂N. Theorem C can be taken as some evidence for this conjecture.

We note that a positive resolution of this conjecture would also imply Minsky's conjecture that the diameter of the *skinning map* is bounded by constants only depending on ∂N, and provide an approach to improving related estimates for the models of [14].

1.1 The Weil–Petersson gradient flow of renormalized volume

One of the main purposes of this paper is to develop the structure theory of the gradient flow V for renormalized volume V_R. From this development, the above results will follow directly. We show that flow provides a powerful new tool to investigate the internal geometry of ends of hyperbolic 3–manifolds.

To give a basic outline of the main ideas of the paper, we begin with a general discussion of gradient flows, which we will then apply to the gradient of renormalized volume. Let f be a smooth function on a noncompact, not necessarily complete, Riemannian manifold X, and assume that

(a) f is bounded below,

(b) the gradient flow of f is defined for all time,

(c) $\|\nabla f\| \le C$,

(d) f has a unique critical point $\bar{x}$,

(e) for all $\epsilon > 0$ there exists an $A > 0$ such that if $d(x, \bar{x}) \ge \epsilon$, then $\|\nabla f\| \ge A$.

By integrating $\|\nabla f\|$ along a distance-minimizing path between points x and x' we immediately see that (c) implies that

$$|f(x) - f(x')| \le C d(x, x').$$

Clearly, we cannot expect a similar lower bound to hold as the level sets of f may have infinite diameter. Instead, we obtain lower bounds when $x' = \bar{x}$, the unique critical point. In particular, let x_t be a flow line of $-\nabla f$ with $x = x_0$. We then have

$$f(x) - f(x_a) = \int_0^a \|\nabla f(x_t)\|^2 \, dt.$$

By (a), $\lim_{a\to\infty} f(x_a)$ exists so as $a \to \infty$, the improper integral is convergent. Therefore there will be an increasing sequence of t_i with $\|\nabla f(t_i)\| \to 0$ so, by (e), the flow line x_t will accumulate on $\bar{x}$. Fix some $\epsilon > 0$ with corresponding $A > 0$ as in (e) and let $I_\epsilon \subset [0,\infty)$ be those values t where $d(x_t,\overline{x}) > \epsilon$. Then for $t \in I_\epsilon$ we have $\|\nabla f(x_t)\| \geq A$ and the length of the path x_t restricted to I_ϵ will be at least $d(x,\overline{x})-\epsilon$. Therefore,

$$\begin{aligned} f(x) - f(\overline{x}) = \int_0^\infty \|\nabla f(x_t)\|^2\,dt &\geq \int_{I_\epsilon} \|\nabla f(x_t)\|^2\,dt \\ &\geq A \int_{I_\epsilon} \|\nabla f(x_t)\|\,dt \geq A(d(x,\overline{x})-\epsilon), \end{aligned}$$

which gives the desired linear lower bound.

Unfortunately, when we replace f with the renormalized volume function V_R, property (e) will not hold (but the others will). To mimic what happens in our generic setting, we let $\overline{X}$ be the metric completion of our Riemannian manifold X and $\mathcal{G} \subset \overline{X}$ a subset. We replace (e) with the following three properties:

(e-1) For all $\epsilon > 0$ there exists a $A > 0$ such that if $d(x,\mathcal{G}) \geq \epsilon$ then $\|\nabla f(x)\| \geq A$.

(e-2) There exists an $n > 0$ such that in any subset of $\mathcal{G}$ with more than n elements there are at least two that a distance δ_0 apart.

(e-3) For every $x_0 \in \mathcal{G}$ there is a path x_t starting at x_0 with $x_t \in X$ for $t > 0$ and $f(x_t) < f(x_0)$.

While the overall structure of the argument will remain the same, some modifications are necessary. First, we need to construct a *surgered flow* x_t where

- $x_0 = x$,
- the function $t \mapsto f(x_t)$ satisfies $f(x_t) < f(x_0)$,
- outside of the ϵ–neighborhood of $\mathcal{G}$, x_t is the gradient flow,
- $x_t \to \overline{x}$ as $t \to \infty$.

To construct x_t we start the gradient flow at x. If it limits to $\overline{x}$ (as we conjecture it will for renormalized volume) then we are done. If not, we limit to some other point in $\mathcal{G}$. We reparametrize so that this happens in finite time and then use (e-3) to restart the flow. If this converges to $\overline{x}$ we stop; if not we repeat. The first three bullets follow directly from this construction.

As before we fix an ϵ and A as in (e-1) and let $I_\epsilon(a) \subset [0,a]$ be those $t \in [0,a]$ where $d(x_t,\mathcal{G}) > \epsilon$. If $L_\epsilon(a)$ is the length of the path $x_{[0,a]}$ restricted to $I_\epsilon(a)$ then the above argument gives

$$f(x) - f(x_a) \geq A L_\epsilon(a).$$

A simple geometric argument, using (e-2), shows that $L_\epsilon(a)$ grows linearly in both the number of points of $\mathcal{G}$ that x_t passes through and in the distance $d(x, x_a)$. In particular, if x_t passes through infinitely many points in $\mathcal{G}$ then $L_\epsilon(a) \to \infty$ as $a \to \infty$ so $f(a) \to -\infty$, contradicting (a). Therefore x_t only passes through finitely many points in $\mathcal{G}$ which implies that the surgered flow converges to the critical point. Therefore if we take the limit of the above inequality we have

$$f(x) - f(\overline{x}) \geq A L_\epsilon(\infty),$$

and as $L_\epsilon(\infty)$ is bounded below by a linear function of $d(x, \overline{x})$, we have our bound.

We now apply this discussion to the renormalized volume function V_R on a relatively acylindrical deformation space $CC(N; S, X)$. Properties (a)–(d) are already known so we will focus on (e-1)–(e-3). In particular, we need to understand when $\|\nabla V_R\|$ is small. By Theorem 1.1 we have that the Weil–Petersson gradient of V_R is given by the *harmonic* Beltrami differential

$$\nabla V_R(M) = \frac{\overline{\phi_M}}{\rho_M},$$

where ρ_M is the area form for the hyperbolic metric on $\partial_c M$ and ϕ_M is the quadratic differential associated to the projective structure on the components of $\partial_c M$ corresponding to S. The norm of ∇V_R is then the L^2–norm of ϕ_M. This L^2–norm is zero exactly when $\phi_M = 0$. As ϕ_M is the Schwarzian derivative of the univalent map uniformizing the components of $\partial_c M$ corresponding to S (see [22]), $\phi_M = 0$ implies that the uniformizing maps are Möbius. It follows that if the norm of ∇V_R is zero then the components of the boundary of the convex core facing S are totally geodesic. In a relatively acylindrical deformation space there is exactly one such manifold (which is why (d) holds) and one might hope that when $\|\phi_M\|_2$ is small we are near this critical point. If this were so, (e) would hold. Unfortunately, it does not. While $\|\phi_M\|_2$ being small will imply that M is near a hyperbolic manifold whose convex core boundary (facing S) is totally geodesic, this manifold may have *rank one cusps*.

To state this more precisely, if $\mathrm{GF}(N; S, X)$ is the space of *geometrically finite* hyperbolic structures on $(N; S, X)$, then the map $M \mapsto \partial_c M$ is a bijection from $\mathrm{GF}(N; S, X)$ to the Weil–Petersson metric completion $\overline{\mathrm{Teich}(S)}$ of Teichmüller space where points in the completion are *noded hyperbolic structures* on S; see [25]. Nodes in the conformal boundary correspond to rank one cusps in the hyperbolic 3–manifold. The triple $(N; S, X)$ determines a subset $\mathcal{G}(N; S, X)$ of $\overline{\mathrm{Teich}(S)}$ where the corresponding hyperbolic structures have totally geodesic boundary facing S. With $\mathcal{G} = \mathcal{G}(N; S, X)$ defined, we can briefly describe how we will verify (e-1)–(e-3).

Property (e-1) is the following theorem and its proof will occupy much of the paper:

Theorem D *For all $\epsilon > 0$, there exists $A = A(\epsilon, S)$ such that if $M \in CC(N; S, X)$ with $\|\phi_M\|_2 \le A$ then there is an $M' \in \mathcal{G}(N; S, X) \subset \mathrm{GF}(N; S, X)$ such that*

$$d_{\mathrm{WP}}(\partial_c M, \partial_c M') \le \epsilon.$$

Property (e-2) follows from Wolpert's strata separation theorem (Theorem 2.2). For a noded surface $Y \in \partial\overline{\mathrm{Teich}(S)}$, we denote the family of curves given by the nodes by τ_Y. Then Wolpert's strata separation theorem implies there is a universal constant $\delta_0 > 0$ such that if $Y_1, Y_2 \in \partial\overline{\mathrm{Teich}(S)}$ with geometric intersection $i(\tau_{Y_1}, \tau_{Y_2}) \ne 0$, then $d_{\mathrm{WP}}(Y_1, Y_2) > \delta_0$. Thus (e-2) holds with $n = 2^{\xi(S)}$, where $\xi(S)$ is the maximal number of disjoint simple closed curves on S as any collection of greater than n noded surfaces in $\partial\overline{\mathrm{Teich}(S)}$ contains two that have intersecting nodes.

Finally property (e-3) follows by unbending the nodes by decreasing the bending angle from π along the nodes to some angle $\theta < \pi$. Such a deformation was constructed by Bonahon and Otal [3]. Using the variational formula for V_R it can be easily shown that V_R satisfies property (e-3) along this path (see Proposition 5.2) as required.

1.2 Constants

A striking feature of Schlenker's proof of the second author's upper bounds for volume is that the constants are very explicit. Unfortunately we lack the same control of constants in our lower bounds as there is one place in the proof, the use of McMullen's contraction theorem for the skinning map, that we fail to control constants explicitly. If we assume, optimistically, that the contraction constant does not depend on the manifold then we can at least understand the asymptotics. With this assumption the multiplicative constant in our lower bound will decay exponentially with exponent of order g^2, where g is the genus. On the other hand, the additive constants will decay to zero even without controlling the contraction constant. This should be compared to work of Aougab, Taylor and Webb [1], who produced an effective lower bound in the quasifuchsian case via the second author's combinatorial methods. Their multiplicative constants decay exponentially with exponent of order $g \log g$, which is better than ours, but their additive constant grows, also of order $g \log g$, rather than decays.

1.3 Questions and conjectures

A central feature of the surgered gradient flow of $-V_R$ on a relatively acylindrical deformation space is that it converges to the unique structure whose convex core has totally geodesic boundary. While in this paper we will focus on relatively acylindrical deformation spaces, the gradient flow is defined on the deformation space of any hyperbolizable 3–manifold as is a surgered flow. We conjecture:

Conjecture 1.2 The surgered gradient flow either converges to a hyperbolic structure whose convex core has totally geodesic boundary or it finds an obstruction to the existence of such a structure. More concretely, either

- N is acylindrical and $M_t \to M_{\mathrm{geod}}$, or
- there is an essential annulus or compressible disk whose boundary has small length in $\partial_c M_t$ for some t.

In fact we expect that the surgeries are unnecessary. Here is a more concrete conjecture when the manifold has incompressible boundary.

Conjecture 1.3 Let N have incompressible boundary. Then for $M \in CC(N)$ the renormalized volume gradient flow M_t starting at M has the property that for any simple closed curve γ on ∂N the geodesic length $\ell_{M_t}(\gamma^*)$ tends to zero if and only if γ lies in the window frame.

See Thurston's paper [33] for the definition of the window of a hyperbolic 3–manifold with incompressible boundary.

In effect, the renormalized volume gradient flow realizes the geometric decomposition of the manifold into pieces by pinching cylinders corresponding to the window boundary, cutting the convex core of the manifold into pared acylindrical pieces with totally geodesic boundary and Fuchsian "windows".

Other questions relate to the internal geometric structure of convex cocompact ends and how the flow relates to their internal structure. To avoid technicalities, for the remainder of this section we will assume that our manifolds are acylindrical.

Let $\mathcal{C}(M, L)$ the collection of simple closed curves on ∂M with geodesic length $\leq L$ in M, and let $\mathcal{F}(M, L)$ be the collection of simple closed curves on ∂M that have length $\leq L$ on some $\partial_c M_t$, where M_t is the gradient flow starting at M.

Question 1.4 Given $L > 0$ does there exist an $L' > 0$ such that

$$\mathcal{F}(M, L') \subset \mathcal{C}(M, L) \quad \text{and} \quad \mathcal{C}(M, L') \subset \mathcal{F}(M, L)?$$

A stronger version of this question is the following.

Question 1.5 Does the flow give a continuous family of bi-Lipschitz embeddings into the initial manifold? In other words, for $s < t$, does the convex core of M_t embed in the convex core of M_s in a bi-Lipschitz manner?

Note that a positive answer to this question would have applications. First, it would imply Thurston's compactness theorem for deformation spaces of acylindrical manifolds. A suitable generalization of this conjecture to the general incompressible case would also imply Thurston's relative compactness theorem in this setting. It would also imply the following conjecture that was mentioned above:

Conjecture 1.6 Let N be an acylindrical 3–manifold. Then for all $M \in CC(N)$ the convex core of M_{geod} has a bi-Lipschitz embedding in M with constants only depending on ∂N.

We note that as gradient flow lines are Weil–Petersson quasigeodesics, relative stability properties established in Brock and Masur [13] for low-genus cases (genus two or lower complexity) for such quasigeodesics would control the behavior of manifolds along the flow M_t when ∂N has genus two. This observation gives an approach to Question 1.5 in such cases. Such stability fails to hold in higher genus cases, so other properties of the flow would be required. The question is reminiscent of similar questions involving the relation of Weil–Petersson geodesics to properties of ends of hyperbolic 3–manifolds and the models of Brock, Canary and Minsky [12].

Acknowledgements We would like to thank MSRI and Yale University for their hospitality while portions of this work were being completed. We also thank Dick Canary, Curt McMullen and Yair Minsky for helpful conversations. We thank the referee for many helpful comments on the initial version. Bridgeman's research was supported by NSF grants DMS-1500545 and DMS-1564410. Brock's research was supported by NSF grants DMS-1608759 and DMS-1849892. Bromberg's research was supported by NSF grants DMS-1509171 and DMS-1906095.

2 Background and notation

In what follows, we fix S to be a closed orientable surface with connected components having genus at least two.

Norms on quadratic differential and metrics on Teichmüller space Let $\Omega^{p,q}(Y)$ be the space of (p,q)–differentials on a Riemann surface Y. Given a *quadratic differential* $\phi \in \Omega^{2,0}(Y)$ and a *Beltrami differential* $\mu \in \Omega^{-1,1}(Y)$, the product $\mu\phi$ is $(1,1)$–differential which can canonically be identified with a 2–form, so we have a pairing

$$\langle \phi, \mu \rangle = \int_Y \mu\phi.$$

In particular, these two spaces are naturally dual.

We also have the subspace $Q(Y) \subset \Omega^{2,0}(Y)$ of *holomorphic quadratic differentials*. This space is important as it is canonically identified with the cotangent space $T^*_Y \operatorname{Teich}(S)$. The tangent space $T_Y \operatorname{Teich}(S)$ is then a quotient of $\Omega^{-1,1}(Y)$. In particular, define

$$N(Y) = \{\mu \in \Omega^{-1,1}(Y) \mid \langle \phi, \mu \rangle = 0 \text{ for all } \phi \in Q(Y)\},$$

and then

$$T_Y \operatorname{Teich}(S) = \Omega^{-1,1}(Y)/N(Y).$$

If ρ_Y is the area form for the hyperbolic metric on Y and $\phi \in \Omega^{2,0}(Y)$, then $|\phi|/\rho_Y$ is also a function, and we define $\|\phi(z)\| = |\phi(z)|/\rho_Y(z)$ to be the pointwise norm. We let $\|\phi\|_p$ be the L^p–norm of this function on Y, again with respect to the hyperbolic area form. Given $\mu \in \Omega^{-1,1}(Y)$ we define the L^q–norm (with $1/p + 1/q = 1$) of the equivalence class $[\mu] \in T_Y \operatorname{Teich}(S)$ by

$$\|[\mu]\|_q = \sup_{\phi \in Q(Y)\setminus\{0\}} \frac{|\langle \phi, \mu \rangle|}{\|\phi\|_p} \le \|\mu\|_q.$$

For $p = 1$ this norm on $T_Y \operatorname{Teich}(S)$ gives the *Teichmüller metric* on $\operatorname{Teich}(S)$ and for $p = 2$ it gives the Weil–Petersson metric. Note that the Teichmüller metric is a Finsler metric while the Weil–Petersson metric is Riemannian, as the L^2–norm on $Q(Y)$ can be given as an inner product. In particular, the L^2–norm on $Q(Y)$ is given by the inner product

$$(\psi, \phi) = \operatorname{Re} \int_Y \psi\bar{\phi}/\rho_Y.$$

From this we see that if $f\colon \mathrm{Teich}(S) \to \mathbb{R}$ is a smooth function then its differential df is an assignment of a holomorphic quadratic differential ϕ_Y to each $Y \in \mathrm{Teich}(S)$. Its Weil–Petersson gradient is the vector field is represented at each Y by a Beltrami differential μ_Y, where for all $\psi \in Q(Y)$ we have

$$(\psi, \phi_Y) = \langle \psi, \mu_Y \rangle.$$

It is a standard fact (and not hard to check directly) that $[\mu_Y]$ is represented by the *harmonic Beltrami differential* $\overline{\phi}_Y/\rho_Y$ and that

$$\|[\mu_Y]\|_2 = \|\overline{\phi}_Y/\rho_Y\|_2 = \|\phi_Y\|_2.$$

Collars We state the collar lemma originally due to Keen [18]. We give it in a form due to Buser [15].

Theorem 2.1 (Buser [15]) *Let Y be a complete hyperbolic surface and γ a simple closed geodesic of length $\ell_\gamma(Y)$. Then the collar $B(\gamma)$ of width*

$$w(\gamma) = \sinh^{-1}\left(\frac{1}{\sinh\left(\frac{1}{2}\ell_\gamma(Y)\right)}\right)$$

is embedded. If $z \in B(\gamma)$, then

$$\sinh(\mathrm{inj}_Y(z)) = \sinh\left(\tfrac{1}{2}\ell_\gamma(Y)\right)\cosh(d(z,\gamma)).$$

Furthermore, for any two disjoint geodesics, the collars are disjoint.

Let $\epsilon_2 = \sinh^{-1}(1)$ be the Margulis constant in dimension 2. If $\ell_\gamma(Y) \le 2\epsilon_2$ then we define the *standard collar* of γ as

$$\{z \in B(\gamma) \mid \mathrm{inj}_Y(z) \le \epsilon_2\}.$$

We note that it follows from the collar lemma (see [15]) that the standard collar consists of all points in Y that lie on a curve of length $\le 2\epsilon_2$ which is homotopic to γ.

For S a finite-type surface, we define $\xi(S)$ to be the maximal number of disjoint simple closed curves in S. For S a surface of genus g and k punctures we have $\xi(S) = 3g - 3 + k$, and for S with connected components S_i then $\xi(S) = \sum_i \xi(S_i)$.

Hyperbolic 3–manifolds Let (N, P) be a pared 3–manifold (see eg [27]) and S a collection of components of $\partial N - P$. Then the triple $(N, P; S)$ is *relatively acylindrical* if no essential cylinder has boundary in S. The acylindricity condition implies that all components of S are incompressible.

A complete hyperbolic 3–manifold M on the interior of N naturally has the structure of a pared 3–manifold. This is simplest to describe when M is geometrically finite and, as this is the only setting we will consider, we stick to this case. Let $\overline{M}$ be the union of M and its conformal boundary. Then there is a paring locus $P \subset \partial N$ such that $\overline{M}$ is homeomorphic to $N - P$. The paring locus P is a collection of annuli and tori. These are the *rank one* and *rank two cusps* of M. In particular, a curve in $M \subset N$ has parabolic holonomy if and only if it is homotopic into P.

Let $\mathrm{MP}(N, P)$ be the space of geometrically finite hyperbolic structures on the interior of N with induced pared manifold structure (N, P). (These are *minimally parabolic* structures on (N, P) — every parabolic is contained in P.) Now fix a conformal structure X on the complement of S in $\partial N - P$ and let $\mathrm{MP}(N, P; S, X) \subset \mathrm{MP}(N, P)$ be those hyperbolic structures with conformal boundary X on the complement of S. Then by the deformation theory of Kleinian groups (see eg [20]) we have the parametrization $\mathrm{MP}(N, P; S, X) \simeq \mathrm{Teich}(S)$. The space $\mathrm{MP}(N, P; S, X)$ is a *quasiconformal deformation space*; any two hyperbolic manifolds in $\mathrm{MP}(N, P; S, X)$ are quasiconformal deformations of each other with the deformation supported on S.

Our results on renormalized volume will only apply to manifolds where P is empty. However, in the course of the proof it will be necessary to consider hyperbolic 3–manifolds with cusps.

Schwarzian derivatives and projective structures Let $f\colon \Delta \to \widehat{\mathbb{C}}$ be a locally univalent map on the unit disk $\Delta \subset \mathbb{C}$. The Schwarzian derivative is the quadratic differential given by

$$Sf(z) = \left(\left(\frac{f''(z)}{f'(z)}\right)' - \frac{1}{2}\left(\frac{f''(z)}{f'(z)}\right)^2\right) dz^2.$$

If f is a Möbius transformation then $Sf = 0$, and in general, Sf measures how much f differs from a Möbius transformation. We also have the composition rule

$$S(f \circ g)(z) = Sf(g(z))g'(z)^2 + Sg(z).$$

Observe that if f is a Möbius transformation then $S(f \circ g) = Sg$, while if g is a Möbius transformation $S(f \circ g)(z) = Sf(g(z))g'(z)^2$.

Let Γ be a Fuchsian group such that $Y = \Delta / \Gamma$. A *projective structure* on Y is given by a locally univalent map $f\colon \Delta \to \widehat{\mathbb{C}}$ (the *developing map*) with a *holonomy representation* $\rho\colon \Gamma \to \mathsf{PSL}_2(\mathbb{C})$ such that for all $\gamma \in \Gamma$ we have

$$f \circ \gamma = \rho(\gamma) \circ f.$$

The composition rule for the Schwarzian implies that Sf descends to a holomorphic quadratic differential in $Q(Y)$.

The Weil–Petersson completion and its stratification While the Teichmüller metric is complete, there are paths with finite length in the Weil–Petersson metric that leave every compact subset of Teichmüller space. Our goal in this section is to describe some of the basic structure of the *completion* of the Weil–Petersson metric. Points in this metric completion are naturally parametrized by families of *Riemann surface with nodes*, namely, a degeneration of a finite-area hyperbolic Riemann surface obtained by collapsing the curves in a multicurve to cusps.

Given a compact surface S, the *complex of curves* $\mathscr{C}(S)$ is a simplicial complex organizing the isotopy classes of simple closed curves on S that do not represent boundary components. To each isotopy class γ we associate a vertex v_γ, and each k–simplex σ is the span of $k+1$ vertices whose associated isotopy classes can be realized disjointly on S.

It is due to Masur [25] that the completion of $\mathrm{Teich}(S)$ with the Weil–Petersson metric is identified with the *augmented Teichmüller space*, obtained by adjoining at infinity the Riemann surfaces with nodes. A point in the completion is given by a choice of the multicurve τ, a (0–skeleton of a) simplex in $\mathscr{C}(S)$, and finite-area hyperbolic structures on the complementary subsurfaces $S \setminus \tau$. The completion is stratified by the simplices of $\mathscr{C}(S)$: the collection of noded Riemann surfaces with nodes determined by a given simplex τ lies in a product of lower-dimensional Teichmüller spaces determined by varying the structures on $S \setminus \tau$. This *stratum* of the completion, $\mathscr{S}_\tau$, inherits a natural metric from the Weil–Petersson metric, which by Masur [25] is isometric to the product of Weil–Petersson metrics on the Teichmüller spaces of the complementary subsurfaces.

The Teichmüller space, with this "augmentation" by its Weil–Petersson completion, naturally descends under the action of the mapping class group to a finite diameter metric on the Deligne–Mumford compactification of the moduli space of Riemann surfaces. If $\overline{\mathrm{Teich}(S)}$ is the completion then we can describe the strata as follows

$$\mathscr{S}_\tau = \{X \in \overline{\mathrm{Teich}(S)} \mid \ell_\gamma(X) = 0 \text{ if and only if } \gamma \in \tau\},$$

where ℓ_γ is the extended length function of γ.

We note that if $\tau_0 \subseteq \tau_1$ are simplices in $\mathscr{C}(S)$, then we have $\mathscr{S}_{\tau_1} \subseteq \overline{\mathscr{S}_{\tau_0}}$.

In his investigation of the geometry of the completion, Wolpert showed the following.

Theorem 2.2 (Wolpert [34, Corollary 22]) *There is a positive constant δ_0 such that either $i(\tau_0, \tau_1) = 0$ and the closures of the strata $\mathscr{S}_{\tau_0}$ and $\mathscr{S}_{\tau_1}$ intersect or $i(\tau_0, \tau_1) > 0$ and*

$$d_{\mathrm{WP}}(\mathscr{S}_{\tau_0}, \mathscr{S}_{\tau_1}) \geq \delta_0.$$

We note that the minimum such δ_0 satisfies $6.57 < \delta_0 < 6.66$; see [7].

3 Hyperbolic 3–manifolds with small Schwarzian derivative

Before proving Theorem D we set some notation. Let $(N, P; S)$ be a relatively acylindrical triple where P is a collection of tori and X a conformal structure on the complement of S in $\partial N - P$. We consider the following:

- τ is a simplex in $\mathcal{C}(S)$.
- P_τ is the union of P and the curves in τ.
- S_τ is the complement of τ in S.

Note that the new triple $(N, P_\tau; S_\tau)$ is still relatively acylindrical and the complement of S_τ in $\partial N - P_\tau$ is homeomorphic to the complement of S in $\partial N - P$. We then have

$$\mathrm{GF}(N, P; S, X) = \bigsqcup_{\tau} \mathrm{MP}(N, P_\tau; S_\tau, X).$$

Thus, $\mathrm{GF}(N, P; S, X)$ is naturally parametrized by the Weil–Petersson completion $\overline{\mathrm{Teich}(S)}$ of Teichmüller space.

We next set:

- If $Y \in \overline{\mathrm{Teich}(S)}$, then M_Y is the hyperbolic manifold in $\mathrm{GF}(N, P; S, X)$ under the above identification $\mathrm{GF}(N, P; S, X) \cong \overline{\mathrm{Teich}(S)}$.
- ϕ_Y is the Schwarzian quadratic differential given by the projective structure on Y induced by M_Y.

We are especially interested in those manifolds in $\mathrm{GF}(N, P; S, X)$ where the boundary of the convex core facing S is totally geodesic. We fix notation for this set:

- Y^τ_{geod} is the unique conformal structure in $\mathrm{Teich}(S_\tau)$ such that the component of the boundary of the convex core of $M_{Y^\tau_{\mathrm{geod}}}$ facing S_τ is totally geodesic.

- $\mathscr{G}(N, P; S, X)$ is the union of the Y^{τ}_{geod}.
- If $\tau = \varnothing$, then we set $Y_{\text{geod}} = Y^{\tau}_{\text{geod}}$ and $M_{\text{geod}} = M_{Y_{\text{geod}}}$.

We have the following elementary observation.

Lemma 3.1 *Let $(N, P; S)$ be a relatively acylindrical triple where P is a collection of tori and X a conformal structure on the complement of S in $\partial N - P$. Then the set $\mathscr{G}(N, P; S, X)$ in $\overline{\text{Teich}(S)}$ is discrete.*

Proof Assume that $Y^{\tau_k}_{\text{geod}} \to Y^{\tau}_{\text{geod}}$ is a convergent sequence in $\mathscr{G}(N, P; S, X)$. Then we can choose an $n > 0$ such that $d_{\text{WP}}(Y^{\tau_k}_{\text{geod}}, Y^{\tau}_{\text{geod}}) < \delta_0/2$ for $k > n$, where δ_0 is the constant in Wolpert's strata separation theorem (Theorem 2.2). By the triangle inequality we also have $d_{\text{WP}}(Y^{\tau_k}_{\text{geod}}, Y^{\tau_l}_{\text{geod}}) < \delta_0$ for $k, l > n$. Thus by Wolpert's strata separation theorem we have $i(\tau_k, \tau_l) = i(\tau_k, \tau) = 0$ for $k, l > n$. This implies that τ_k can be only a finite number of possibilities for $k > n$ and therefore $\mathscr{G}(N, P; S, X)$ is discrete. $\square$

We will also be interested in the manifold obtained by *drilling* the curves in τ from the interior of N. We set notation here:

- Set $W \cong \partial N \times [0, 1]$ to be a collar neighborhood of ∂N with $\partial_0 W = \partial N \times \{0\}$ the component of the boundary lying in $\text{int}(N)$.
- Set $\tau_0 = \tau \times \{0\}$ to be copies of τ isotoped into $\text{int}(N)$, lying on $\partial_0 W$.
- Let $\widehat{N}$ be the compact 3–manifold obtained removing open tubular neighborhoods $\mathscr{N}(\tau_0)$ of τ_0.
- Note that $\partial \widehat{N}$ is the union of ∂N and a torus for each component of τ_0. Let $\widehat{P}$ be the union of P and the new tori in $\partial \widehat{N}$ so there is a natural homeomorphism from $\partial N - P$ to $\partial \widehat{N} - \widehat{P}$.

There is an inclusion $\iota\colon \widehat{N} \hookrightarrow N$ that restricts to a homeomorphism from $\partial \widehat{N} - \widehat{P}$ to $\partial N - P$. Therefore $\text{MP}(\widehat{N}, \widehat{P}; S, X)$ is also parametrized by $\text{Teich}(S)$.

- Given $Y \in \text{Teich}(S)$, $\widehat{M}_Y \in \text{MP}(\widehat{N}, \widehat{P}; S, X)$ is the hyperbolic manifold such that ι extends to a conformal map between the conformal boundary of $\widehat{M}_Y$ and M_Y.
- $\widehat{\phi}_Y$ is the Schwarzian quadratic differential for the projective structure on Y induced by $\widehat{M}_Y$.

There is a natural embedding

$$j\colon N \to \hat{N}$$

obtained by including the submanifold $N \setminus \operatorname{int}(W \cup \mathcal{N}(\tau_0)) \hookrightarrow N$ such that the composition $\iota \circ j$ is isotopic to the identity and j is a homeomorphism from $\partial N - (P \cup S)$ to $\partial \hat{N} - (\hat{P} \cup S)$.

For every hyperbolic manifold in $\mathrm{MP}(\hat{N}, \hat{P}; S, X)$ this embedding induces a cover that lies in $\mathrm{MP}(N, P_\tau; S_\tau, X)$. That is, there is an induced map

$$j^*\colon \mathrm{MP}(\hat{N}, \hat{P}; S, X) \to \mathrm{MP}(N, P_\tau; S_\tau, X)$$

between the deformation spaces and we set

$$M_{\hat{Y}} = j^*(\hat{M}_Y).$$

Outline of the proof of Theorem D If $\|\phi_Y\|_\infty$ is small the proof is straightforward: Thurston's skinning map is a map from $\mathrm{MP}(N, P; S, X)$ to itself that has a fixed point at the totally geodesic structure. By a theorem of McMullen the skinning map is contracting and therefore we obtain a bound on the distance from Y to Y_{geod} if we can bound distance between Y and its first skinning iterate. When $\|\phi_Y\|_\infty$ is small, a classical result of Ahlfors and Weill bounds this initial distance.

A key element of our investigation involves understanding the behavior of the L^∞–norm when the L^2–norm is small. In particular, the pointwise norm of ϕ_Y may be large in the thin parts of Y which we will need to pinch to nodes. There are several steps to the proof:

- We choose τ to be the simplex of short curves on Y. A version of the *drilling theorem* bounds the L^2–norm of $\phi_Y - \hat{\phi}_Y$ in terms of the length of τ. We use this to bound the pointwise norm of $\hat{\phi}_Y$ outside of the standard collars of τ.
- Using the above bullet and a modification of some classical arguments, this bounds $\|\phi_{\hat{Y}}\|_\infty$. We are then in position to use McMullen's contraction theorem to bound the distance between $\hat{Y}$ and Y^τ_{geod}.
- We also have that $Y - \tau$ conformally embeds in $\hat{Y}$, which implies that Y and $\hat{Y}$ are close in the Weil–Petersson completion. Together, this and the previous bullet point imply the theorem.

3.1 Choosing the curves to drill

As we noted in the outline, a bound in $\|\phi_Y\|_2$ does not give a bound on $\|\phi_Y\|_\infty$. However, we have the following bound on the pointwise norm that depends on the injectivity radius. For Y a hyperbolic surface and $z \in Y$ we define $\mathrm{inj}_Y(z)$ to be the injectivity radius of z in the hyperbolic metric on Y. For simplicity, we define the truncated injectivity radius by $\mathrm{inj}_Y^-(z) = \min\{\mathrm{inj}_Y(z), \epsilon_2\}$, where $\epsilon_2 = \sinh^{-1}(1)$ is the Margulis constant in dimension 2.

Proposition 3.2 (Bridgeman and Wu [8]) *Let $\phi \in Q(Y)$ then*

$$\|\phi(z)\| \leq \frac{\|\phi\|_2}{\sqrt{\mathrm{inj}_Y^-(z)}}.$$

As a first step we show that after an appropriate choice for τ, we can obtain a pointwise bound on $\widehat{\phi}_Y$ outside of the standard collars of τ. For this we will need the following bound on the L^2–norm.

Theorem 3.3 (Bridgeman and Bromberg [5]) *There exist constants $c_{\mathrm{drill}}, \ell_{\mathrm{drill}} > 0$ with $\ell_{\mathrm{drill}} < 1$ such that the following holds. Given $Y \in \mathrm{Teich}(S)$ and a simplex τ in $\mathcal{C}(S)$ such that $\ell_\beta(Y) \leq \ell_{\mathrm{drill}}$ for all $\beta \in \tau$, we have*

$$\|\phi_Y - \widehat{\phi}_Y\|_2 \leq c_{\mathrm{drill}} \sqrt{\ell_\tau(Y)},$$

where $\ell_\beta(Y)$ is the length of β in Y.

Fixing a universal constant We first prove that we can choose the simplex τ such that $\|\widehat{\phi}_Y(z)\|$ is small for $z \in Y$ in the complement of the standard collars of τ.

Theorem 3.4 *Assume that $Y \in \mathrm{Teich}(S)$ with $\|\phi_Y\|_2^{2/(2\xi(S)+3)} \leq \ell_{\mathrm{drill}}$. There exists an $\ell = \ell(Y) > 0$ with*

$$\ell \leq \|\phi_Y\|_2^{2/(2\xi(S)+3)}$$

such that the following holds. Let τ be the simplex in $\mathcal{C}(S)$ of all curves with length $\leq \ell$. Then for $z \in Y$ in the complement of the standard collars of τ,

$$\|\widehat{\phi}_Y(z)\| \leq C_0 \sqrt{\xi(S)} \|\phi_Y\|_2^{2/(2\xi(S)+3)}.$$

for $C_0 = \sqrt{2}(c_{\mathrm{drill}} + 1)$.

Proof Let $\Lambda = \|\phi_Y\|_2^{2/(2\xi(S)+3)} \le \ell_{\mathrm{drill}} < 1$ and let $\ell_k = \Lambda^{2k+1}$. As $\Lambda < 2\epsilon_2$, there are at most $\xi(S)$ curves of length $\le \Lambda$ so there must be some integer k with $0 \le k \le \xi(S)$ such that Y has no curves of length in the interval $(\ell_{k+1}, \ell_k]$. Let $\ell = \ell_{k+1} \le \ell_0 = \|\phi_Y\|_2^{2/(2\xi(S)+3)}$ and let τ be the simplex in $\mathscr{C}(S)$ of all curves of length $\le \ell$ on Y.

By Theorem 3.3 we have

$$\|\phi_Y - \widehat{\phi}_Y\|_2 \le c_{\mathrm{drill}}\sqrt{\ell_\tau(Y)}.$$

As $\ell_\tau(Y) \le \xi(S)\Lambda^{2k+3}$, we have

$$\|\widehat{\phi}_Y\|_2 \le \|\phi_Y\|_2 + \|\phi_Y - \widehat{\phi}_Y\|_2 \le \Lambda^{\xi(S)+\frac{3}{2}} + c_{\mathrm{drill}}\sqrt{\xi(S)}\Lambda^{k+\frac{3}{2}}.$$

As Y contains no curves of length in the interval $(\ell_{k+1}, \ell_k]$ every point in the complement of the standard collars of τ has injectivity radius $> \ell_k/2 = \Lambda^{2k+1}/2$. Therefore if $z \in Y$ is in the complement of the standard collars of $\mathcal{C}$, then by Proposition 3.2

$$\begin{aligned}\|\widehat{\phi}_Y(z)\| \le \frac{\|\widehat{\phi}_Y\|_2}{\sqrt{\ell_k/2}} &\le \frac{\Lambda^{\xi(S)+\frac{3}{2}} + c_{\mathrm{drill}}\sqrt{\xi(S)}\Lambda^{k+\frac{3}{2}}}{\sqrt{\Lambda^{2k+1}/2}}\\ &\le \sqrt{2}(\Lambda + c_{\mathrm{drill}}\sqrt{\xi(S)}\Lambda)\\ &\le \sqrt{2}(1 + c_{\mathrm{drill}}\sqrt{\xi(S)})\Lambda\\ &\le C_0\sqrt{\xi(S)}\Lambda,\end{aligned}$$

where $C_0 = \sqrt{2}(1 + c_{\mathrm{drill}})$ is a universal constant. □

We can now prove:

Theorem 3.5 *If* $Y \in \mathrm{Teich}(S)$ *with*

$$\|\phi_Y\|_2^{2/(2\xi(S)+3)} \le \min\{\ell_{\mathrm{drill}}, 2\sinh^{-1}(\tfrac{1}{2})\},$$

then there is a simplex $\tau \in \mathscr{C}(S)$ *and a* $\widehat{Y} \in \mathrm{Teich}(S_\tau) \subseteq \overline{\mathrm{Teich}(S)}$ *such that*

(a) $d_{\mathrm{WP}}(Y, \widehat{Y}) \le \dfrac{2\pi}{\sqrt{\sinh^{-1}(\frac{1}{2})}}\sqrt{\xi(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)}$,

(b) $\|\phi_{\widehat{Y}}\|_\infty \le C_1\sqrt{\xi(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)}$,

where $C_1 = 9\sqrt{2}(C_0 + 1)$.

Proof of Theorem 3.5(a) Let τ be the simplex given by Theorem 3.4 and let $\hat{Y} \in \operatorname{Teich}(S_\tau)$ be the surface with $j^*(\hat{M}_Y) = M_{\hat{Y}}$. To obtain the bound on $d_{\mathrm{WP}}(Y, \hat{Y})$ we will apply Proposition A.1, and to do this we need to show that certain covers of Y embed in $\hat{Y}$. To set notation, let $\hat{\Gamma}_Y$ be a Kleinian group such that $\hat{M}_Y = \mathbb{H}^3/\hat{\Gamma}_Y$. Then $M_{\hat{Y}} = \mathbb{H}^3/\Gamma_{\hat{Y}}$, where $\Gamma_{\hat{Y}} \subset \hat{\Gamma}_Y$ is a subgroup.

We consider the domains of discontinuity of these two groups. First, note that as $\Gamma_{\hat{Y}}$ is a subgroup of $\hat{\Gamma}_Y$, the domain of discontinuity of $\Gamma_{\hat{Y}}$ contains the domain of discontinuity of $\hat{\Gamma}_Y$. More precisely, if Γ is the subgroup of $\Gamma_{\hat{Y}}$ that fixes a component Ω of the domain discontinuity of $\Gamma_{\hat{Y}}$ then the subgroup Γ will be the fundamental group of the one of the components of the boundary of the pared manifold $M_{\hat{Y}}$. Under the inclusion $j : M_{\hat{Y}} \hookrightarrow \hat{M}_Y$, boundary components of the pared manifold $M_{\hat{Y}}$ will be homotopic to embeddings into components of the pared manifold $\hat{M}_Y$. As Γ corresponds to the fundamental group of a component of the boundary of $M_{\hat{Y}}$, this implies that there will be a subgroup $\hat{\Gamma}$ of $\hat{\Gamma}_Y$, corresponding to the fundamental group of a component of the boundary of $\hat{M}_Y$, with Γ a subgroup of $\hat{\Gamma}$. Then $\hat{\Gamma}$ will fix a component $\hat{\Omega}$ of the domain of discontinuity of $\hat{\Gamma}_Y$. As Γ is a subgroup of $\hat{\Gamma}$ it will also fix $\hat{\Omega}$ and therefore $\hat{\Omega} \subset \Omega$. We finally note that if Ω/Γ is a component of X, as j restricted to X is a homeomorphism, we have $\hat{\Gamma} = \Gamma$ and $\hat{\Omega} = \Omega$.

Fix a component $\hat{W}$ of $\hat{Y}$ and let $\Omega_{\hat{W}}$ be a component of the domain discontinuity that covers $\hat{W}$. Let $\Gamma_{\hat{W}} \subset \Gamma_{\hat{Y}}$ be the subgroup that fixes $\Omega_{\hat{W}}$. Then $\hat{W} = \Omega_{\hat{W}}/\Gamma_{\hat{W}}$. By the above, there is a component W of Y, a component $\hat{\Omega}_W$, and a subgroup $\hat{\Gamma}_W$ of $\hat{\Gamma}_Y$ with

- $W = \hat{\Omega}_W/\hat{\Gamma}_W$,
- $\Gamma_{\hat{W}} \subset \hat{\Gamma}_W$,
- $\hat{\Omega}_W \subset \Omega_{\hat{W}}$.

As $\Gamma_{\hat{W}}$ also fixes $\hat{\Omega}_W$, the quotient $\check{W} = \hat{\Omega}_W/\Gamma_{\hat{W}}$ embeds in $\hat{W} = \Omega_{\hat{W}}/\Gamma_{\hat{W}}$, where $\check{W}$ is the cover of W corresponding to the (topological) inclusion $\hat{W} \hookrightarrow W$.

Let $\check{Y}$ be the union of the covers $\check{W}$ of (the components of) Y obtained by letting $\hat{W}$ vary over all components of $\hat{Y}$. Then $\hat{Y}$ embeds in $\check{Y}$, and by assumption, we have

$$\|\phi_Y\|_2^{2/(2\xi(S)+3)} \le 2\sinh^{-1}\left(\tfrac{1}{2}\right).$$

Therefore, by Theorem 3.4, for each $\beta \in \tau$,

$$\ell_\beta(Y) \le 2\sinh^{-1}\left(\tfrac{1}{2}\right)$$

so we can apply Proposition A.1 to get

$$d_{\mathrm{WP}}(Y,\widehat{Y}) \le \frac{2\pi}{\sqrt{\sinh^{-1}\left(\frac{1}{2}\right)}}\sqrt{\ell_\tau(Y)} \le \frac{2\pi}{\sqrt{\sinh^{-1}\left(\frac{1}{2}\right)}}\sqrt{\xi(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)}. \qquad \square$$

To obtain our bound on $\|\phi_{\widehat{Y}}\|_\infty$ we will first need the following generalization of the Kraus–Nehari bound on the norm of the Schwarzian.

Lemma 3.6 *Let $f\colon \Delta \to \Delta$ be univalent and assume that for $z \in \Delta$ the image $f(\Delta)$ contains a hyperbolic disk of radius r centered at $f(z)$. Then $\|Sf(z)\| \le \frac{3}{2}\operatorname{sech}\left(\frac{1}{2}r\right)$.*

Proof The proof is a refinement of the classical proof of the Kraus–Nehari theorem.

Assume that $z = f(z) = 0$. By applying the Schwarz lemma to the restriction of f^{-1} to the hyperbolic disk of radius r, we see that $|f'(0)| \ge \tanh\left(\frac{1}{2}r\right)$. If we let $g(z) = f'(0)/f(1/z)$ we have the expansion

$$g(z) = z + \sum_{n=0}^{\infty} b_n z^{-n}.$$

Note that the domain of g is $\{z \in \widehat{\mathbb{C}} \mid |z| > 1\}$ and that $|g(z)| > \tanh\left(\frac{1}{2}r\right)$ for z in the domain. As in the proof of Nehari's theorem we can also calculate to see that $Sf(0) = -6b_1$. As the conformal factor for the area form of the hyperbolic metric on Δ at $z = 0$ is 4, we obtain $\|Sf(0)\| = \frac{3}{2}|b_1|$. Let C_ρ be the circle of radius ρ centered at 0 with $\rho > 1$. Then the Euclidean area m_ρ in $\mathbb{C}$ bounded by $g(C_\rho)$ is

$$m_\rho = \pi\rho^2 - \pi\sum_{n=1}^{\infty} n|b_n|^2\rho^{-2n}.$$

Since, for all $\rho > 1$, C_ρ will contain the disk of radius $\tanh\left(\frac{1}{2}r\right)$ centered at 0 we have that $m_\rho > \pi\tanh^2\left(\frac{1}{2}r\right)$ and by letting $\rho \to 1$ we have

$$\pi\tanh^2\left(\tfrac{1}{2}r\right) \le \pi - \pi\sum_{n=1}^{\infty} n|b_n|^2 \le \pi - \pi|b_1|^2.$$

The estimate follows. $\square$

Proof of Theorem 3.5(b) Choose ϵ such that

$$\|\phi_Y\|_2^{2/(2\xi(S)+3)} = 2\epsilon,$$

and note that $2\epsilon \le \epsilon_2$.

We use the same setup as in (a). As we have there, $\Omega_{\widehat{W}}$ is component of the domain of discontinuity $\Gamma_{\widehat{Y}}$ covering a component $\widehat{W} \subset \widehat{Y}$ of the conformal boundary of $M_{\widehat{Y}}$. We need to bound the Schwarzian of the uniformizing map $f_{\widehat{W}}: \Delta \to \Omega_{\widehat{W}}$. If $f_W: \Delta \to \widehat{\Omega}_Y$ is the map uniformizing $\widehat{\Omega}_W \subset \Omega_{\widehat{W}}$ the we can factor f_W through a map $g: \Delta \to \Delta$ such that $f_W = f_{\widehat{W}} \circ g$. Here g is the lift of the embedding $\check{W} \hookrightarrow \widehat{W}$ described above. To control the Schwarzian of $f_{\widehat{W}}$ we need to apply Lemma 3.6 to g and combine the bound there with the given bounds on the Schwarzian of f_W.

Let $\widehat{W}^{\epsilon_2}$ and $\widehat{W}^{\epsilon}$ be the complements of the ϵ_2– and ϵ–cuspidal thin parts of $\widehat{W}$, respectively. By the Schwarz lemma the embedding $\check{W} \hookrightarrow \widehat{W}$ is a contraction from the complete hyperbolic metric on $\check{W}$ (which is lifted from $W \subset Y$) to the complete hyperbolic metric on $\widehat{W}$. The peripheral curves in $\check{W}$ will map to the cuspidal curves in $\widehat{W}$. In W these curves are in τ and therefore have length in W (and therefore in $\check{W}$) that is $\leq 2\epsilon$. This implies that the image of embedding of $\check{W}$ in $\widehat{W}$ will contain $\widehat{W}^{\epsilon}$. At the level of universal covers this implies that if $z \in \Delta$ such that $f_{\widehat{W}}(z)$ is mapped into $\widehat{W}^{\epsilon}$ in the quotient $\Omega_{\widehat{W}} / \Gamma_{\widehat{W}}$ then z is in the image of g.

By [2, Lemma 4.5] the norm of a quadratic differential achieves its maximum in the complement of the standard neighborhood of the cusps. Therefore to bound $\|\phi_{\widehat{Y}}\|_\infty$ it suffices to bound $\|\phi_{\widehat{W}}(z)\|$ for $z \in \widehat{W}^{\epsilon_2}$.

After fixing a $z \in \widehat{W}^{\epsilon_2}$ it will be convenient to normalize our uniformizing maps so that $g(0) = 0$ and 0 maps to z under the quotient maps to $\widehat{W}$ and Y. Then

$$\|\phi_{\widehat{Y}}(z)\| = 4|Sf_{\widehat{W}}(0)| \quad \text{and} \quad \|\widehat{\phi}_Y(z)\| = 4|Sf_W(0)|.$$

By the composition rule for Schwarzian derivatives we have

$$Sf_W(0) = Sf_{\widehat{W}}(g(0))g'(0)^2 + Sg(0),$$

and therefore (assuming that $g(0) = 0$)

$$\|\phi_{\widehat{Y}}(z)\| = \|Sf_{\widehat{W}}(0)\| \leq \frac{\|Sf_W(0)\| + \|Sg(0)\|}{|g'(0)|^2}.$$

We now need to bound the individual terms on the right.

As $\widehat{W}^{\epsilon_2}$ is in the complement of the standard collars of τ in Y, by Theorem 3.4

$$\|Sf_W(0)\| = \|\phi_Y(z)\| \leq 2C_0\sqrt{\xi(S)}\epsilon.$$

We would like to apply Lemma 3.6 to bound $\|Sg(0)\|$ but to do so we need to bound from below the distance from 0 to $\Delta \backslash g(\Delta)$ in the hyperbolic metric on Δ. This distance is bounded below by the distance from $\widehat{W}^{\epsilon_2}$ to $\widehat{W}\backslash\breve{W}$ in the hyperbolic metric on $\widehat{W}$, and this distance in turn is bounded below by the distance from $\widehat{W}^{\epsilon_2}$ to $\widehat{W}\backslash\widehat{W}^{\epsilon}$ since $\widehat{W}^{\epsilon}$ is contained in $\breve{W}$. A simple calculation shows that if r is the distance from $\partial\widehat{W}^{\epsilon_2}$ to $\partial\widehat{W}^{\epsilon}$, then

$$e^r > \frac{\sinh\left(\frac{1}{2}\epsilon_2\right)}{\sinh\left(\frac{1}{2}\epsilon\right)} > \frac{\epsilon_2}{\epsilon} \geq 2.$$

The hyperbolic disk of radius r centered at 0 will be contained in $g(\Delta)$ and Lemma 3.6 plus the above bound implies that

$$\|Sg(z)\| \leq \tfrac{3}{2}\operatorname{sech}\left(\tfrac{1}{2}r\right) < 3e^{-\frac{1}{2}r} < 3\sqrt{\frac{\epsilon}{\epsilon_2}}.$$

Finally we need to bound from below $|g'(0)|$. As in the proof of Lemma 3.6 we have $|g'(0)| \geq \tanh\left(\frac{1}{2}r\right)$, and given our above bound on r this becomes

$$|g'(0)| \geq \tanh\left(\tfrac{1}{2}r\right) \geq \frac{1-\epsilon/\epsilon_2}{1+\epsilon/\epsilon_2} \geq \tfrac{1}{3}.$$

Combining our estimates we have

$$\|Sf_{\widehat{W}}(0)\| \leq 9\left(2C_0\sqrt{\xi(S)}\epsilon + 3\sqrt{\frac{\epsilon}{\epsilon_2}}\right) \leq 9\sqrt{2}(C_0+1)\sqrt{\xi(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)}.$$

Therefore we let $C_1 = 9\sqrt{2}(C_0+1)$, and the result follows. □

3.2 Bounds on iteration of the skinning map

Let $(N, P; S)$ be a relatively acylindrical triple. For $Y \in \operatorname{Teich}(S) \cong \operatorname{MP}(N, P; S, X)$ we need to show that if $\|\phi_Y\|_\infty$ is small, then $d_{\mathrm{WP}}(Y, Y_{\mathrm{geod}})$ is small. When (N, P) is acylindrical the proof is a straightforward application of a classical bound of Ahlfors and Weill plus McMullen's contraction theorem for the skinning map. However, in the relatively acylindrical case we will need a slight extension of McMullen's original statement.

The *skinning map*

$$\sigma\colon \operatorname{MP}(N, P; S, X) \simeq \operatorname{Teich}(S) \to \operatorname{Teich}(S)$$

is defined as follows: for each $Y \in \operatorname{Teich}(S)$, the cover of $M_Y \in \operatorname{MP}(N, P; S, X)$ associated to the subgroup $\pi_1(Y) \subset \pi_1(M_Y)$ under inclusion will be quasifuchsian. (If Y is disconnected then the cover will also be a finite collection of a quasifuchsian manifolds.) For each connected component of ∂M_Y, one component of the conformal boundary restricts to a homeomorphism to Y under the covering projection. The other component will be $\sigma(Y)$, the image of the skinning map for that component. Note that $Z \in \mathrm{Teich}(S)$ is in $\mathcal{G}(N, P; S, X)$ if and only if Z is a fixed point for σ.

The skinning map is a smooth map and we will be interested in bounding its derivative so that we can apply the contraction mapping principle. The estimate we need from McMullen essentially works as written in [26] but there are a few differences in the relative case, which we highlight. Given $Y \in \mathrm{Teich}(S)$ let Γ be the Kleinian group that uniformizes $M_Y \in \mathrm{MP}(N, P; S, X)$ and let Ω be the domain of discontinuity of Γ. If the pair (N, P) was acylindrical, then every component of Ω would be a Jordan domain and the stabilizer of every component would be a quasifuchsian group. Furthermore if D_0 and D_1 are distinct components of Ω then either their closures are disjoint and the intersection of their stabilizers is trivial, or the intersection is a point and the intersection of their stabilizers is an infinite cyclic group generated by a parabolic. In the relatively acylindrical case this will not hold. However, if we let Ω_Y be those components of Ω that cover Y then these properties do hold for the components in Ω_Y. The second key point is that a tangent vector of $\mathrm{MP}(N, P; S, X)$ is represented by a Γ–invariant Beltrami differential μ that is supported on Ω_Y. With these two observations one sees that McMullen's proof in the acylindrical case extends to the relatively acylindrical case:

Theorem 3.7 (McMullen [26, Theorem 6.1 and Corollary 6.2]) *If $(N, P; S, X)$ is relatively acylindrical, then for $Y \in \mathrm{Teich}(S)$,*

$$\|d\sigma_Y\|_\infty \le \lambda(S) < 1,$$

where $\lambda(S)$ depends only on the topology of S.

The contraction mapping principle implies that $\sigma^n(Y) \to Z$ with $\sigma(Z) = Z$ and

$$d_{\mathrm{Teich}}(Y, Z) \le \frac{d_{\mathrm{Teich}}(Y, \sigma(Y))}{1 - \lambda(S)}.$$

To complete the proof of Theorem 3.9 we need to bound $d(Y, \sigma(Y))$. This is a direct consequence of the Ahlfors–Weill quasiconformal reflection theorem:

Theorem 3.8 (Ahlfors and Weill [23, Theorem 5.1]) *Let $Y \in \mathrm{Teich}(S)$ and ϕ_Y be the associated quadratic differential on Y. If $\|\phi_Y\|_\infty < \frac{1}{2}$ then*

$$d_{\mathrm{Teich}}(Y, \sigma(Y)) \le \tfrac{1}{2} \log \frac{1 + 2\|\phi_Y\|_\infty}{1 - 2\|\phi_Y\|_\infty}.$$

If $\|\phi_Y\|_\infty \leq \frac{1}{3}$, then an easy estimate of the right-hand side gives

$$d_{\mathrm{Teich}}(Y, \sigma(Y)) \leq 3\|\phi_Y\|_\infty,$$

and therefore

$$d_{\mathrm{Teich}}(Y, Z) \leq \frac{3}{1-\lambda(S)}\|\phi_Y\|_\infty.$$

By a result of Linch [24], $d_{\mathrm{WP}} \leq \sqrt{\mathrm{area}(S)}\, d_{\mathrm{Teich}}$ and we have the following result.

Theorem 3.9 *Let $(N, P; S)$ be relatively acylindrical. Then for all $Y \in \mathrm{Teich}(S)$ with $\|\phi_Y\|_\infty \leq \frac{1}{3}$ we have*

$$\frac{d_{\mathrm{WP}}(Y, Z)}{\sqrt{\mathrm{area}(Y)}} \leq d_{\mathrm{Teich}}(Y, Y_{\mathrm{geod}}) \leq \frac{3\|\phi_Y\|_\infty}{1-\lambda(S)},$$

where $\lambda(S)$ is the contraction constant from Theorem 3.7.

Remark McMullen's proof is not effective and this is the one place in our proof where we don't control the growth rate of the constants in terms of genus. However, we have made some effort to isolate this from the constants that we do control.

3.3 Proof of Theorem D

We now put together the results above. We first restate Theorem D, but here we carefully control the constants.

Theorem 3.10 *There are a universal constants K_0 and ϵ_0 such that if*

$$A(\epsilon, S) = \left(\frac{K_0\epsilon(1-\lambda(S))}{\xi(S)}\right)^{2\xi(S)+3}$$

and $Y \in \mathrm{Teich}(S)$ with $\|\phi_Y\|_2 \leq A(\epsilon, S)$ and $\epsilon \leq \epsilon_0$ then there exists $Y^\tau_{\mathrm{geod}} \in \mathcal{G}$ with $d_{\mathrm{WP}}(Y, Y^\tau_{\mathrm{geod}}) \leq \epsilon$.

Proof By Theorem 3.5, there are universal constants $\ell_{\mathrm{drill}}, C_1 > 0$ such that if $\|\phi_Y\|_2^{2/(2\xi(S)+3)} \leq \ell_{\mathrm{drill}}$ then there is a simplex τ in $\mathcal{C}(S)$ such that after drilling curves $\mathcal{C}$,

$$\|\phi_{\hat{Y}}\|_\infty \leq C_1\sqrt{\xi(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)},$$

$$d_{\mathrm{WP}}(Y, \hat{Y}) \leq \frac{2\pi}{\sqrt{\sinh^{-1}\left(\frac{1}{2}\right)}}\sqrt{\xi(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)}.$$

Assuming that $\|\phi_{\widehat{Y}}\|_\infty \le \frac{1}{3}$ we can apply Theorem 3.9 to $(N_\tau, P_\tau; S_\tau)$ to see that

$$d_{\mathrm{WP}}(\widehat{Y}, Y^\tau_{\mathrm{geod}}) \le \frac{3\sqrt{\operatorname{area}(\widehat{Y})}}{1-\lambda(S)}\|\phi_{\widehat{Y}}\|_\infty \le \frac{2\sqrt{3\pi}C_1\xi(S)}{1-\lambda(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)},$$

since $\operatorname{area}(\widehat{Y}) = \operatorname{area}(Y) = \frac{4}{3}\pi\xi(S)$. Then by the triangle inequality and the fact that $C_1 > 1$, we have

$$d_{\mathrm{WP}}(Y, Y^\tau_{\mathrm{geod}}) \le d_{\mathrm{WP}}(Y, \widehat{Y}) + d_{\mathrm{WP}}(\widehat{Y}, Y^\tau_{\mathrm{geod}}) \le \frac{4\sqrt{3\pi}C_1\xi(S)}{1-\lambda(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)}.$$

We let $K_0 = 1/(4\sqrt{3\pi}C_1)$. Recounting our progress, if

$$\|\phi_Y\|_2 \le A(\epsilon, S) = \left(\frac{K_0\epsilon(1-\lambda(S))}{\xi(S)}\right)^{2\xi(S)+3},$$

we have

$$d_{\mathrm{WP}}(Y, Y^\tau_{\mathrm{geod}}) \le \epsilon,$$

assuming that $\|\phi_Y\|_2^{2/(2\xi(S)+3)} < \ell_{\mathrm{drill}}$ and $\|\phi_{\widehat{Y}}\|_\infty \le \frac{1}{3}$. However, if we let

$$\epsilon_0 = \min\left\{\frac{\sqrt{\ell_{\mathrm{drill}}}}{K_0}, 4\sqrt{\frac{\pi}{3}}\right\}$$

and $\epsilon < \epsilon_0$ then

$$\|\phi_Y\|_2^{2/(2\xi(S)+3)} \le \left(\frac{K_0\epsilon(1-\lambda(S))}{\xi(S)}\right)^2 \le (K_0\epsilon)^2 < \ell_{\mathrm{drill}}$$

and

$$\begin{aligned}\|\phi_{\widehat{Y}}\|_\infty &\le C_1\sqrt{\xi(S)}\|\phi_Y\|_2^{1/(2\xi(S)+3)}\\ &\le C_1\sqrt{\xi(S)}\frac{K_0\epsilon(1-\lambda(S))}{\xi(S)}\\ &\le C_1K_0\epsilon = \frac{\epsilon}{4\sqrt{3\pi}} \le \frac{1}{3}.\end{aligned}$$

This completes the proof. □

Hyperbolic manifolds with cylinders and compression disks We conclude this section with a discussion of where we use the relative acylindricity of $(N, P; S)$. For simplicity, in this discussion we will assume that both P and S are empty.

The first problem that can occur is in Theorem 3.4 and its application. In particular, it can happen that two or more curves in τ may be homotopic in N or even homotopically trivial in N if N has compressible boundary. In this case, the manifold $\widehat{N}$ will not be hyperbolizable. If N has incompressible boundary, this problem can be corrected by only removing a single curve from N for each homotopy class (in N) of curves in τ. With this change, Theorem 3.4 we still hold but we cannot define the embedding of N in $\widehat{N}$ and therefore cannot carry through the proof of Theorem 3.5.

If none of the curves in τ are homotopic in N then the proofs up to and including Theorem 3.5 go through. However, if the pared manifold (N, P_τ) is not acylindrical then Theorem 3.7, McMullen's contraction theorem, will fail. In fact, the deformation space $\mathrm{MP}(N, P_\tau)$ contains a hyperbolic structure whose convex core boundary is totally geodesic if and only if (N, P_τ) is acylindrical or is a pared I–bundle.

We expect that the only problem that can occur is the first one. We have the following conjecture.

Conjecture 3.11 Let M be a convex cocompact hyperbolic 3–manifold with ϕ the Schwarzian quadratic differential for the projective boundary of M. If $\|\phi\|_2$ is small, then either:

- There exists a geometrically finite structure M' on N with totally geodesic convex core boundary, and $d_{\mathrm{WP}}(\partial_c M, \partial_c M')$ is small.
- There are two or more short curves on $\partial_c M$ that are homotopic in M.

In particular, if no two curves in τ are homotopic in M, we expect that (N, P_τ) is an acylindrical pair even when N itself is not acylindrical.

4 W–volume and renormalized volume

Given a convex submanifold N with smooth boundary such that $N \hookrightarrow M$ is a homotopy equivalence, the *W–volume* of N is defined to be

$$W(N) = \mathrm{vol}(N) - \frac{1}{2}\int_{\partial N} H\, dA,$$

where H is the mean curvature[1] of ∂N.

The W–volume has many nice analytic properties that make it a useful tool for studying hyperbolic manifolds. We let N_t be the t–neighborhood of N. The nearest point

[1]This differs from the formula in [21] as we define $H = \mathrm{Tr}(B)/2$ rather than $H = \mathrm{Tr}(B)$, where B is the shape operator.

retraction from M to each N_t extends to a diffeomorphism from $\partial_c M$ to ∂N_t and using this retract we pull back the induced metrics on ∂N_t to metrics I_t on $\partial_c M$. Then

$$I^*(x,y) = \lim_{t\to\infty} \frac{1}{\cosh^2(t)} I_t(x,y)$$

is a well-defined metric in the conformal class of $\partial_c M$ and is called the *metric at infinity*.

For $N \subset M$ we will denote by ρ_N the metric at infinity on $\partial_c M$. The W–volume has the following properties.

Proposition 4.1 (Krasnov and Schlenker [21]) *Let $N \subset M$ be a compact, convex submanifold of a convex cocompact hyperbolic 3–manifold M and let N_t be the t–neighborhood of N. Then:*

(1) *The metric ρ_N is in the conformal class of $\partial_c M$.*

(2) $\rho_{N_t} = e^{2t}\rho_N$.

(3) $W(N_t) = W(N) - t\pi\chi(\partial N)$.

Furthermore, if ρ is any smooth conformal metric on $\partial_c M$ then for t sufficiently large there exists a convex submanifold $X_t \subset M$ with $\rho_{X_t} = e^{2t}\rho$.

Using this proposition, the W–volume of any smooth conformal metric ρ on $\partial_c M$ is defined by

$$W(\rho) = W(N_t(\rho)) + t\pi\chi(\partial M)$$

for t sufficiently large. The proposition above implies that $W(\rho)$ doesn't depend on the choice of t. With this setup we can now define the *renormalized volume* V_R by setting

$$V_R(M) = W(\rho_M),$$

where ρ_M is the unique hyperbolic metric on $\partial_c M$.

Convex cores Perhaps the most natural convex submanifold of a convex cocompact hyperbolic 3–manifold M is the *convex core* $C(M)$. The boundary of the convex core is not in general smooth, so we cannot use the previous definition to define the W–volume of $C(M)$. However, there is a natural way to extend W–volume to this setting (see the discussion in [4]) and for the convex core we have

$$W(C(M)) = V_C(M) - \tfrac{1}{4}L(\beta_M),$$

where β_M is the *bending lamination* of the boundary of the convex core and $L(\beta_M)$ is its length (as a measured lamination). The convex core also induces a natural metric at infinity, called the *projective metric* (so called as Thurston gave a definition that is intrinsic to the induced projective structure on $\partial_c M$). We will be interested in a hybrid metric that is the hyperbolic metric on some components of $\partial_c M$ and the projective metric on the others. We have the following:

Proposition 4.2 *Let M be a convex cocompact hyperbolic 3–manifold and suppose that $\partial_c M = X \sqcup Y$, a disjoint union of connected components of $\partial_c M$. Let σ be the hyperbolic metric on X and the projective metric on Y. Let β_Y be the bending lamination of the components of the boundary of $C(M)$ that faces Y. Then*

$$W(\sigma) - \tfrac{1}{4}L(\beta_Y) \le V_R(M) \le W(\sigma).$$

In particular, if $Y = \partial N$, we have

$$V_C(M) - \tfrac{1}{2}L(\beta_M) \le V_R(M) \le V_C(M) - \tfrac{1}{4}L(\beta_M).$$

By the definition of the W–volume of the convex core, the two statements are equivalent for the case $X = \varnothing$, and this case was proven in [4, Theorem 3.7]. Furthermore, the proof trivially extends to the relative case above.

5 The variational formula

Recall that if $(N; S)$ is a pair such that each component of S is incompressible in N then $\mathrm{MP}(N; S, X)$ is parametrized by $\mathrm{Teich}(S)$ and therefore we can view renormalized volume as a function

$$V_R\colon \mathrm{Teich}(S) \to \mathbb{R}.$$

We recall the variational formula:

Theorem 1.1 *Given $Y \in \mathrm{Teich}(S)$ and $\mu \in T_Y \mathrm{Teich}(S)$, we have*

$$dV_R(\mu) = \mathrm{Re} \int_{\partial_c M_Y} \phi_Y \mu.$$

Therefore the Weil–Petersson gradient of V_R has norm $\|\phi_Y\|_2$. By the classical bound of Kraus–Nehari for the Schwarzian of univalent functions, we have that $\|\phi_Y\|_\infty \le \tfrac{3}{2}$. As a corollary we have:

Corollary 5.1 *The Weil–Petersson norm of the gradient of V_R is bounded by*

$$\tfrac{3}{2}\sqrt{\mathrm{area}(Y)} = \sqrt{3\pi n(S)}.$$

In particular, V_R is Lipschitz with respect to the Weil–Petersson metric, and therefore extends to a continuous function on the Weil–Petersson completion.

Note if S is not incompressible in N then we cannot apply the Kraus–Nehari theorem to bound the norm of the gradient and in fact there is no upper bound of the gradient in this setting.

We now assume that $(N;S)$ is relatively acylindrical and recall that $\mathcal{G} = \mathcal{G}(N;S,X)$ is the collection of $Y \in \overline{\mathrm{Teich}(S)}$ such that the component of the boundary convex core of M_Y facing Y is totally geodesic.

Proposition 5.2 *Given τ a nonempty simplex in $\mathcal{C}(S)$, let Y^τ_{geod} be the unique surface in $\mathcal{G} \cap \mathrm{Teich}(S_\tau)$. Then for $t > 0$ there is a one-parameter family $Y_t \in \mathrm{Teich}(S)$ with $Y_t \to Y^\tau_{\mathrm{geod}}$ as $t \to 0$ with $V_R(Y_t) < V_R(Y^\tau_{\mathrm{geod}})$.*

Proof By a construction of Bonahon and Otal [3] there exists a one-parameter family $M_\theta \in \mathrm{MP}(N;S,X)$ where the bending lamination β_θ of the components of the convex core facing S have support τ and bending angle θ. In the parametrization $\mathrm{MP}(N;S,X) \cong \mathrm{Teich}(S)$, the manifolds M_θ correspond to $Z_\theta \in \mathrm{Teich}(S)$. We also let σ_θ be the hybrid metric that is the projective metric on Z_θ and the hyperbolic metric on X. Let ϕ_θ be the Schwarzian quadratic differential on Z_θ.

As part of the construction, Bonahon and Otal show that M_{Z_θ} converges to $M_{Y^\tau_{\mathrm{geod}}}$ in the *algebraic topology* on $\mathrm{GF}(N,S;X)$. Unfortunately what we need is that $Z_\theta \to Y^\tau_{\mathrm{geod}}$ in $\overline{\mathrm{Teich}(S)} \cong \mathrm{GF}(N,S;X)$, where the topology is the metric topology of the Weil–Petersson completion. These two topologies are not homeomorphic. While the convergence we need could be proven using the notion of strong convergence of Kleinian groups and techniques well-known to experts, we will instead give a proof more in line with the methods from this paper.

We first note that from the construction it follows that $L(\beta_\theta) \to 0$ as $\theta \to \pi$. In [6] it is shown that

$$\|\phi_\theta\|_2 \le \tfrac{5}{2}\sqrt{L(\beta_\theta)},$$

and therefore we also have $\|\phi_\theta\|_2 \to 0$ as $\theta \to \pi$. Theorem 3.10 then implies that Z_θ accumulates on $\mathcal{G}$. As $\mathcal{G}$ is discrete (see Lemma 3.1), Z_θ must limit to a unique point.

It also follows from construction that the length of a curve γ on Z_θ limits to zero if and only if γ is in τ so any limit for Z_θ will be in the strata for τ. Together this implies that $Z_\theta \to Y^\tau_{\mathrm{geod}}$.

By Corollary 5.1, V_R extends to a continuous function on $\overline{\mathrm{Teich}(S)}$. Combining this with Proposition 4.2 and the fact the $L(\beta_\theta) \to 0$ we have

$$\lim_{\theta\to\pi} W(\sigma_\theta) = \lim_{\theta\to\pi} V_R(M_\theta) = V_R(Y^\tau_{\mathrm{geod}}).$$

We will show that

$$V_R(M_\theta) \le W(\sigma_\theta) < V_R(Y^\tau_{\mathrm{geod}}),$$

which will give the result.

For this we use the variational formula

$$\frac{d}{d\theta} W(\sigma_\theta) = \tfrac{1}{4}\big(\ell(\theta) - \theta\ell'(\theta)\big),$$

where $\ell(\theta)$ is the sum of the length of the curves in τ on in M_θ. If $X = \varnothing$ then by the Schläfli formula

$$\frac{d}{d\theta} V_C(M_\theta) = \tfrac{1}{2}\ell(\theta),$$

and the variational formula follows from differentiating the formula for W–volume of the convex core and the noting that $L(\beta_\theta) = \theta\ell(\theta)$. In general, if ρ_t is a family of metrics on ∂N then the variation of W–volume will have a term for each component of the boundary and if $\tilde{\rho}_t$ is a another family of metrics that agrees with ρ_t on a component S of ∂N then the term for both variations on S will be the same. In our case σ_θ is the hyperbolic metric on X for all θ, so the variation of W–volume on X is zero. On Z_θ, σ_θ is the projective metric so on Y the variation is the same as the variation of the W–volume of the convex core. This gives the variational formula.

We can now complete the proof. By Choi and Series [16], $\ell'(\theta) < 0$, which implies that $W(\sigma_\theta) < V_R(Y^\tau_{\mathrm{geod}})$. We can also see this directly by integrating to get

$$V_R(Y^\tau_{\mathrm{geod}}) - W(\sigma_T) = \frac{1}{4}\int_T^\pi \ell(\theta)\,d\theta + \tfrac{1}{8}T\ell(T) > 0.$$

We then define Y_t by reparametrizing Z_θ via an orientation-reversing homeomorphism from $(0,\infty)$ to $(0,\pi)$. Thus we have $V_R(Y_t) < V_R(Y^\tau_{\mathrm{geod}})$, as required. □

6 Lower bounds on renormalized volume

We begin with a geometric lemma. We note that a *geodesic metric space* is a metric space (X, d), where the distance between two points is attained by the length of a path between the points.

Lemma 6.1 *Let Z be a collection of points in a geodesic metric space (X, d) such that for any collection of $n+1$ points in Z there are two that are at least a distance δ apart. Let*

$$\alpha\colon [0,1] \to X$$

be a rectifiable path and let $L_\epsilon(\alpha)$ be the length of the path that is disjoint from the ϵ–neighborhood of Z. Then for $\epsilon < \delta/2n$,

$$L_\epsilon(\alpha) \geq \frac{\delta - 2n\epsilon}{\delta}\big(d(\alpha(0), \alpha(1)) - 2n\epsilon\big).$$

Proof For each $z \in Z$, let

$$U_z = \{t \in [0,1] \mid d(\alpha(t), z) < \epsilon\}$$

and let U be the union of the U_z. Note that for any $t \in [0,1]$ there are at most N points $z \in Z$ such that $\mathcal{N}_\epsilon(z)$ intersects the $(\delta - 2\epsilon)/2$–neighborhood of $\alpha(t)$ and therefore there is a neighborhood of t that intersects at most N of the U_z. As $[0,1]$ is compact this implies that there are finitely many $z \in Z$ with $U_z \neq \varnothing$.

We claim we that we can find $z_1, \dots, z_m$ in Z and

$$0 = t_0^+ \leq t_1^- < t_1^+ \leq t_2^- < \cdots \leq t_m^- < t_m^+ \leq 1 = t_{m+1}^-$$

such that

- $t_i^- \in \bar{U}_{z_i}$,
- $t_i^+ = \sup U_{z_i}$,
- $\alpha([t_{i-1}^+, t_i^-])$ is disjoint from $\mathcal{N}_\epsilon(Z)$.

We assume that the first i points and values have been chosen and then find z_{i+1} and $t_{i+1}^\pm$. Let t_{i+1}^- be the infimum of $(t_i^+, 1] \cap U$. As there are finitely many nonempty U_z, there must be some $z \in Z$ with t_{i+1}^- the infimum of $(t_i^+, 1] \cap U_z$. We let $z_{i+1} = z$ and $t_{i+1}^+ = \sup U_z$. This process terminates (and $m = i$) when either $(t_i^+, 1] \cap U = \varnothing$ or $t_i^+ = 1$.

Note that this implies that $d(\alpha(t_i^-),\alpha(t_i^+)) \le 2\epsilon$ and

$$\sum d(\alpha(t_{i-1}^+),\alpha(t_i^-)) \le L_\epsilon(\alpha).$$

Therefore,

$$d(\alpha(0),\alpha(1)) \le \sum d(\alpha(t_i^-),\alpha(t_i^+)) + \sum d(\alpha(t_{i-1}^+),\alpha(t_i^-)) \le 2m\epsilon + L_\epsilon(\alpha).$$

We need to show that $2m\epsilon$ is only a controlled portion of $d(\alpha(0),\alpha(1))$. For this we choose a nonnegative integer k such that $kn < m \le (k+1)n$. Then we let j_1 be the smallest index such that there exists an $i_1 < j_1$ with $d(z_{i_1}, z_{j_1}) \ge \delta$. Note that $j_1 \le n+1$. Then, as above,

$$\delta - 2\epsilon \le d(\alpha(t_{i_1}^+),\alpha(t_{j_1}^-)) \le 2(n-1)\epsilon + L_\epsilon(\alpha|_{[t_{i_1}^+, t_{j_1}^-]}).$$

Repeating this argument we get i_ℓ and j_ℓ for $\ell = 1, \dots, k$, where $j_{\ell-1} \le i_\ell < j_\ell$, $j_\ell - j_{\ell-1} \le N$ and

$$\delta - 2\epsilon \le d(\alpha(t_{i_\ell}^+),\alpha(t_{j_\ell}^-)) \le 2(n-1)\epsilon + L_\epsilon(\alpha|_{[t_{i_\ell}^+, t_{j_\ell}^-]}).$$

Summing these inequalities and rearranging we get

$$k \le \frac{L_\epsilon(\alpha)}{\delta - 2n\epsilon}.$$

As $m \le (k+1)n$ our previous bound on $d(\alpha(0),\alpha(1))$ becomes

$$d(\alpha(0),\alpha(1)) \le 2(k+1)n\epsilon + L_\epsilon(\alpha).$$

Combining the two inequalities and rearranging gives the result. □

Lemma 6.2 *Assume that $0 < \epsilon \le \epsilon_0$ and let Y_t be a path on $\overline{\mathrm{Teich}(S)}$ such that on $E = \{t \mid d_{\mathrm{WP}}(Y_t, \mathcal{G}) > \epsilon\}$ the path is smooth and the tangent vector is the Weil–Petersson gradient of $-V_{\boldsymbol{R}}$, and for $[u,v]$ a connected component of the path Y_t in E^c we have $V_{\boldsymbol{R}}(Y_v) \le V_{\boldsymbol{R}}(Y_u)$. Then*

$$V_{\boldsymbol{R}}(Y_a) - V_{\boldsymbol{R}}(Y_b) \ge A(\epsilon, S)\frac{\delta_0 - 2^{\xi(S)+1}\epsilon}{\delta_0}(d_{\mathrm{WP}}(Y_a, Y_b) - 2^{\xi(S)+1}\epsilon).$$

Proof We have that E is a collection $\mathcal{I}$ of open intervals. By assumption, for $t \in E$ the tangent vector $\dot{Y}_t$ of Y_t is the Weil–Petersson gradient of $-V_{\boldsymbol{R}}$, so by Theorem 1.1,

$$\|\dot{Y}_t\|_{\mathrm{WP}} = \|\phi_{Y_t}\|_2.$$

By Theorem 3.10 we also have that for $t \in E$,

$$\|\phi_{Y_t}\|_2 \ge A(\epsilon, S).$$

Again applying the variational formula, Theorem 1.1, to an interval (s,t) in $\mathcal{I}$, we have

$$V_{\boldsymbol{R}}(Y_s)-V_{\boldsymbol{R}}(Y_t)=\int_s^t \|\phi_{Y_t}\|_2^2\,dt \geq \int_s^t A(\epsilon,S)\|\phi_{Y_t}\|_2\,dt = A(\epsilon,S)L(Y_{(s,t)}),$$

where $L(Y_{(s,t)})$ is the length of the path from s to t. For any interval $[u,v]$ in E^c, by assumption we have $V_{\boldsymbol{R}}(Y_u)-V_{\boldsymbol{R}}(Y_v)>0$. Therefore we have

$$V_{\boldsymbol{R}}(Y_a)-V_{\boldsymbol{R}}(Y_b)\geq \sum_{(s,t)\in\mathcal{I}} V_{\boldsymbol{R}}(Y_s)-V_{\boldsymbol{R}}(S_t),$$

and therefore

$$V_{\boldsymbol{R}}(Y_a)-V_{\boldsymbol{R}}(Y_b)\geq A(\epsilon,S)L(Y_{\mathcal{I}}),$$

where

$$L(Y_{\mathcal{I}})=\sum_{(s,t)\in\mathcal{I}} L(Y_{(s,t)}).$$

For any collection of $2^{\xi(S)}+1$ simplices in $\mathcal{C}(S)$ there must be at least two that contain intersecting curves. Therefore by Theorem 2.2 for any collection of $2^{\xi(S)}+1$ points in $\mathcal{G}=\mathcal{G}(N;S,X)$ there are at least two that are a distance δ_0 apart in the Weil–Petersson metric on $\overline{\mathrm{Teich}(S)}$ and we can apply Lemma 6.1 with $Z=\mathcal{G}$ the set of points and $n=2^{\xi(S)}$. Noting that $L_\epsilon(Y_{[a,b]})=L(Y_{\mathcal{I}})$ by Lemma 6.1, we have

$$L(Y_{\mathcal{I}})\geq \frac{\delta_0-2^{\xi(S)+1}\epsilon}{\delta_0}\big(d_{\mathrm{WP}}(Y_a,Y_b)-2^{\xi(S)+1}\epsilon\big).$$

Combining this with our above bound on the differences between renormalized volumes gives the result. □

Convergence in the Weil–Petersson completion

Proposition 6.3 *Let Y_t be a flow line of the Weil–Petersson gradient flow of $-V_{\boldsymbol{R}}$. Then Y_t converges in $\overline{\mathrm{Teich}(S)}$ to a $\widehat{Y}\in\mathcal{G}$.*

Proof By Lemma 6.2 for every positive distance $d>0$ there is a $v>0$ such that if $d_{\mathrm{WP}}(Y_s,Y_t)\geq d$ then $V_{\boldsymbol{R}}(Y_s)-V_{\boldsymbol{R}}(Y_t)\geq v$. Renormalized volume is bounded below (and is in fact nonnegative) and therefore $V_{\boldsymbol{R}}(Y_t)$ converges as $t\to\infty$. In particular there exists a $T>0$ such that if $s,t>T$ then $V_{\boldsymbol{R}}(Y_s)-V_{\boldsymbol{R}}(Y_t)<v$ and $d_{\mathrm{WP}}(Y_s,Y_t)<d$. It follows that Y_t converges in $\overline{\mathrm{Teich}(S)}$ as $t\to\infty$.

The lower bound on renormalized volume also implies that the integral

$$\int_0^\infty \|\phi_{Y_t}\|_2^2 \, dt < \infty.$$

Therefore we can find a sequence t_i with $\|\phi_{Y_{t_i}}\|_2 \to 0$ as $i \to \infty$. Theorem 3.10 then implies that any accumulation point of the sequence will lie in $\mathcal{G}$. As we have just seen that the entire path converges, this implies that the limit of Y_t as $t \to \infty$ lies in $\mathcal{G}$. $\square$

The surgered flow

Proposition 6.4 *Fix $\epsilon > 0$. For all $Y \in \mathrm{Teich}(S)$ there exists a path Y_t in $\overline{\mathrm{Teich}(S)}$ with $Y = Y_0$ such that:*

- *On $\{t \mid d_{\mathrm{WP}}(Y_t, \mathcal{G}) > \epsilon\}$, the path is smooth and the tangent vector is the Weil–Petersson gradient of $-V_R$.*
- *If $a < b$ and $[a, b]$ is a connected component of the set $\{t \mid d_{\mathrm{WP}}(Y_t, \mathcal{G}) \le \epsilon\}$, then $V_R(Y_b) < V_R(Y_a)$.*
- *$Y_t \to Y_{\mathrm{geod}}$ as $t \to \infty$.*

Proof We claim there exists an integer $k \ge 0$ such that for $i = 0, \dots, k$ there are a family of paths Y_t^i and simplices $\tau_0, \dots, \tau_k$ in $\mathcal{C}(S)$ such that

- $Y = Y_0^i$,
- Y_t^i passes through $Y_{\mathrm{geod}}^{\tau_0}, \dots, Y_{\mathrm{geod}}^{\tau_{i-1}}$,
- $V_R(Y_{\mathrm{geod}}^{\tau_{j-1}}) < V_R(Y_{\mathrm{geod}}^{\tau_j})$ for $j = 1, \dots i-1$,
- if $d_{\mathrm{WP}}(Y_t^i, \mathcal{G}) > \epsilon$ the path is smooth and the tangent vector $\dot{Y}_t^i$ is the Weil–Petersson gradient of $-V_R$,
- $Y_t^i \to Y_{\mathrm{geod}}^{\tau_i}$ as $t \to \infty$ and $\tau_k = \varnothing$.

We start by letting Y_t^0 be the flow line of the Weil–Petersson gradient of $-V_R$ with $Y_0^0 = Y$. By Proposition 6.3, there is a simplex τ_0 in $\mathcal{C}(S)$ such that Y_t converges to some $Y_{\mathrm{geod}}^{\tau_0} \in \mathcal{G}$, where τ_0 are the nodes of $Y_{\mathrm{geod}}^{\tau_0}$.

Now assume $Y_t^0, \dots, Y_t^i$ and $\tau_0, \dots, \tau_i$ have been chosen. If $\tau_i = \varnothing$ then $k = i$ and we are done. If not, we form Y_t^{i+1} as follows. As $Y_t^i \to \tau_i$ there exists a t_0 such that if $t > t_0$ then $d_{\mathrm{WP}}(Y_t, Y_{\mathrm{geod}}^{\tau_i}) < \epsilon/2$. By Proposition 5.2, there is a path Z_t with $Z_0 = Y_{\mathrm{geod}}^{\tau_i}$, $Z_t \in \mathrm{Teich}(S)$ and $V_R(Z_t) < V_R(Y_{\mathrm{geod}}^{\tau_i})$. We can then choose t_1 such that if $0 < t < t_1$ then $d_{\mathrm{WP}}(Y_{\mathrm{geod}}^{\tau_i}, Z_t) < \epsilon/2$. We then define Y_t^{i+1} by

- $Y_t^{i+1} = Y_t^i$ if $t \le t_0$,

- $Y^{i+1}_{[t_0,t_0+1)}$ is a reparametrization of $Y^i_{[t_0,\infty)}$,
- $Y^{i+1}_t = Z_{t-t_0-1}$ if $t \in [t_0+1, t_0+t_1+1]$,
- for $t \ge t_0+t_1+1$, Y^{i+1}_t is a flow line of the Weil–Petersson gradient of $-V_R$.

For large t, Y^{i+1}_t is a gradient flow line, so once again by Proposition 6.3, we have that $Y^{i+1}_t \to Y^{\tau_{i+1}}_{\text{geod}} \in \mathcal{G}$, where curves in the simplex τ_{i+1} are the nodes of $Y^{\tau_{i+1}}_{\text{geod}}$.

We now show that the process terminates. Observe that $V_R(Y^{\tau_{i+1}}_{\text{geod}}) < V_R(Y^{\tau_i}_{\text{geod}})$ as the path Y^{i+1}_t passes through $Y^{\tau_i}_{\text{geod}}$, $V_R(Y^{i+1}_t)$ is decreasing, and $V_R(Y^{i+1}_t) \to V_R(Y^{\tau_{i+1}}_{\text{geod}})$ as $t \to \infty$ by Corollary 5.1. Thus all of the τ_i are distinct and $V_R(Y^{\tau_i}_{\text{geod}})$ is decreasing in i.

The flows Y^i_t satisfy the conditions of Lemma 6.2 so there exists a $v = v(\epsilon, \delta_0) > 0$ such that if $d_{\text{WP}}(Y^i_a, Y^i_b) \ge \delta_0$ then $V_R(Y^i_a) - V_R(Y^i_b) \ge v$. As we noted above, for any collection of $2^{\xi(S)}+1$ simplices in $\mathcal{C}(S)$ there will be at least two that contain intersecting curves. Therefore for any $i \ge 0$ there exist $j < \ell$ in $\{i, \dots, i+2^{\xi(S)}\}$ such that τ_j and τ_ℓ contain intersecting curves. By Theorem 2.2 we then have $d_{\text{WP}}(Y^{\tau_j}_{\text{geod}}, Y^{\tau_\ell}_{\text{geod}}) \ge \delta_0$. As $Y^{i+2^{\xi(S)}+1}_t$ passes through τ_j and τ_ℓ, in that order (with possibly $i = j$ or $\ell = i+2^{\xi(S)}$), we have
$$V_R(Y^{\tau_i}_{\text{geod}}) - V_R(Y^{\tau_{i+2^{\xi(S)}}}_{\text{geod}}) \ge V_R(Y^{\tau_j}_{\text{geod}}) - V_R(Y^{\tau_\ell}_{\text{geod}}) \ge v.$$

Therefore, if the paths are defined up to i with $2^{\xi(S)}m \le i \le 2^{\xi(S)}(m+1)$, we have
$$V_R(Y) - V_R(Y^{\tau_i}_{\text{geod}}) \ge V_R(Y^{\tau_0}_{\text{geod}}) - V_R(Y^{\tau_i}_{\text{geod}}) \ge mv.$$

As $V_R \ge 0$ this implies that
$$i \le 2^{\xi(S)}\left(\frac{V_R(Y)}{v} + 1\right).$$

Therefore the process must terminate. □

We now use the above to give a new proof of the following theorem of Storm.

Corollary 6.5 (Storm [30; 31]) *Let N be a compact hyperbolizable acylindrical 3–manifold without torus boundary components. Then V_C has a unique minimum at the structure $M_{\text{geod}} \in CC(N)$ with totally geodesic convex core boundary.*

The minimality of M_{geod} was the main result in [30] and the uniqueness is a corollary of the main result in [31], which considers the general case of N with incompressible boundary.

Proof Let $Y \neq Y_{\mathrm{geod}}$. Using surgered flow, we have the path Y_t with $Y_t \in \overline{\mathrm{Teich}(\partial N)}$ from Y to Y_{geod} with $V_R(M_Y) > V_R(M_{\mathrm{geod}})$. Therefore

$$V_C(M_Y) \geq V_R(M_Y) > V_R(M_{\mathrm{geod}}) = V_C(M_{\mathrm{geod}}).$$

Thus V_C has unique minimum at M_{geod}. □

In the course of the proof we have shown that the unique minimum of V_R also occurs at M_{geod}. In the relatively acylindrical case, we no longer have $V_C(M_{\mathrm{geod}}) = V_R(M_{\mathrm{geod}})$, but otherwise the above proof goes through to give the following more general version of Storm's theorem for renormalized volume.

Corollary 6.6 *Let $(N; S)$ be a compact hyperbolizable relatively acylindrical 3–manifold without torus boundary components. Then V_R has a unique minimum at the structure $M_{\mathrm{geod}} \in CC(N; S, X)$ with totally geodesic convex core boundary facing S.*

In [4] we proved that Corollaries 6.5 and 6.6 are equivalent. Here we are directly proving both statements. A version of Corollary 6.6 was also proved by Pallete [28] using different methods.

Also applying Lemma 6.2 to the surgered flow path gives:

Theorem 6.7 *For all $\epsilon \leq \epsilon_0$,*

$$V_R(Y) - V_R(Y_{\mathrm{geod}}) \geq A(\epsilon, S) \frac{\delta_0 - 2^{\xi(S)+1}\epsilon}{\delta_0} (d_{\mathrm{WP}}(Y, Y_{\mathrm{geod}}) - 2^{\xi(S)+1}\epsilon).$$

Theorem A then follows from the above by choosing $\epsilon = \min(\epsilon_0, \delta_0/2^{\xi(S)+2})$ and letting

$$A(S) = \tfrac{1}{2} A(\epsilon, S) \quad \text{and} \quad \delta = \tfrac{1}{2}\delta_0.$$

We also recall Schlenker's upper bounds. His argument was originally for quasifuchsian manifolds, but as we will see it holds whenever $(N; S)$ has incompressible boundary.

Theorem 6.8 *Let $(N; S)$ have incompressible boundary. Then*

$$|V_R(Y) - V_R(Y')| \leq 3\sqrt{\frac{\pi}{2}|\chi(\partial N)|}\, d_{\mathrm{WP}}(Y, Y').$$

Proof As noted in Corollary 5.1 the norm of the Weil–Petersson gradient of V_R is bounded above by $\frac{3}{2}\sqrt{\mathrm{area}(Y)} = 3\sqrt{(\pi/2)|\chi(S)|}$. Integrating this bound along a Weil–Petersson geodesic segment from Y to Y' gives the result. □

We can now use the above to prove Theorem B, which we now restate.

Theorem B *Let S be a closed surface of genus $g \geq 2$. Then*

$$A(S)(d_{\mathrm{WP}}(X,Y)-\delta) \leq V_C(Q(X,Y)) \leq 3\sqrt{\frac{\pi}{2}|\chi(S)|}\, d_{\mathrm{WP}}(X,Y) + 6\pi|\chi(S)|.$$

Proof If $N = S\times[0,1]$ then a *Bers slice* is the deformation space $CC(N; S\times\{0\}, X)$, where X is a fixed conformal structure on S. Manifolds in this deformation space are *quasifuchsian* and the manifold $M_Y \in CC(N; S\times\{0\}, X)$ in our general notation is usually referred to as $Q(X,Y)$.

We apply Theorem A to this case. Then $Q(X,X)$ is the Fuchsian manifold so $Y_{\mathrm{geod}} = X$ and $V_{\boldsymbol{R}}(Y_{\mathrm{geod}}) = 0$. Therefore we have

$$A(S)(d_{\mathrm{WP}}(X,Y)-\delta) \leq V_{\boldsymbol{R}}(Q(X,Y)).$$

Combining this lower bound with the bound of Schlenker [29, Theorem 1.2], we have

$$A(S)(d_{\mathrm{WP}}(X,Y)-\delta) \leq V_{\boldsymbol{R}}(Q(X,Y)) \leq 3\sqrt{\frac{\pi}{2}|\chi(S)|}\, d_{\mathrm{WP}}(X,Y).$$

By [4], for any convex cocompact M,

$$V_{\boldsymbol{R}}(M) + \tfrac{1}{4}L(\beta_M) \leq V_C(M) \leq V_{\boldsymbol{R}}(M) + \tfrac{1}{2}L(\beta_M).$$

Also for ∂N incompressible $L(\beta_M) \leq 6\pi|\chi(\partial N)|$; see [4]. The result follows. □

Theorem C follows identically as in the proof of Theorem B above.

Appendix A Weil–Petersson estimate

We recall that the Margulis constant in two dimensions is $\epsilon_2 = \sinh^{-1}(1)$. In this section we prove the following proposition:

Proposition A.1 *Let τ be a simplex in $\mathcal{C}(S)$ and $Y \in \mathrm{Teich}(S)$ a hyperbolic surface such that $\ell_\beta(Y) \leq \ell_0$ for each curve $\beta \in \tau$, where $0 < \ell_0 < 2\epsilon_2$. Let $\widehat{Y} \in \mathrm{Teich}(S_\tau)$ be such that the cover $\check{Z}$ of Y associated to $S\backslash\tau$ conformally embeds in $\widehat{Y}$. Then*

$$d_{\mathrm{WP}}(Y,\widehat{Y}) \leq 2\pi\sqrt{\frac{2\sinh\left(\frac{1}{2}\ell_0\right)}{\ell_0\left(1-\sinh\left(\frac{1}{2}\ell_0\right)\right)}}\sqrt{\ell_\tau(Y)}.$$

We will use the following criteria for convergence in the Weil–Petersson completion. Let τ be a simplex in $\mathcal{C}(S)$ and $\widehat{Y}$ a surface in $\mathrm{Teich}(S_\tau)$. Then a sequence $Y_i \in \mathrm{Teich}(S)$ converges to $\widehat{Y}$ in $\overline{\mathrm{Teich}(S)}$ if for all simple closed curves γ with $i(\gamma,\tau)=0$ we have $\ell_\gamma(Y_i) \to \ell_\gamma(\widehat{Y})$. In particular the length of the curves in τ must converge to zero. We will use the following lemma to verify this criteria.

Lemma A.2 *Let $R \subset S$ be a proper, essential, nonannular subsurface of a finite-type surface S. Let R_i and S_i be conformal structures on R and S, respectively, such that there is a conformal embedding $R_i \hookrightarrow S_i$ in the homotopy class of $R \hookrightarrow S$. If $\ell_{\partial R}(R_i) \to 0$, then for all simple closed curves γ on R we have*

$$\lim_{i\to\infty} \ell_\gamma(R_i) = \lim_{i\to\infty} \ell_\gamma(S_i),$$

where the lengths are measured on the completed hyperbolic metrics on the respective conformal structures.

Proof Let R_i^γ and S_i^γ be the annular covers of R_i and S_i corresponding to the curve γ. Then there is a conformal embedding $R_i^\gamma \hookrightarrow S_i^\gamma$ that is a homotopy equivalence. Therefore

$$\frac{\pi}{\ell_\gamma(R_i)} = m(R_i^\gamma) \le m(S_i^\gamma) = \frac{\pi}{\ell_\gamma(S_i)},$$

where $m(\cdot)$ is the modulus of the annulus.

To get a bound in the other direction we let D_i be the distance, in the S_i–metric, from the geodesic representative of γ in S_i to the complement of R_i and denote the D_i–neighborhood of the geodesic core of S_i^γ by $S_i^\gamma(D_i)$. Then $S_i^\gamma(D_i)$ will be contained in R_i and it follows that

$$m(S_i^\gamma(D_i)) = \frac{\pi - \epsilon_i}{\ell_\gamma(S_i)} \le m(R_i^\gamma),$$

where ϵ_i only depends on D_i and $\epsilon_i \to 0$ as $D_i \to \infty$. To finish the proof we need to show that $D_i \to \infty$.

Let $C(R_i)$ be the convex core of R_i and assume that each component of the boundary of $C(R_i)$ has length $< 2\epsilon_2$. Then each component of the boundary of $C(R_i)$ will lie in the standard collar of the associated geodesic in S_i. As the length of the boundary curves of $C(R_i)$ limits to zero, the depth of these curves in the standard S_i–collars will limit to infinity. In particular, the distance of any point in the R–component of the complement of the S_i–collars from the complement of the R_i will also limit to infinity. As the geodesic representative of γ in S_i will be in this complementary region we have that $D_i \to \infty$, as desired. □

Let A be a conformal annulus with finite modulus $m(A)$. Then A can be realized as the quotient of the strip

$$S = \{z \in \mathbb{C} \mid 0 < \operatorname{Im} z < \pi\}$$

by the translation

$$z \mapsto z + \frac{\pi}{m(A)}.$$

Define Beltrami differentials μ_A^t and μ_A^h so that their lifts to S are $\tilde{\mu}_A^t = 1$ and $\tilde{\mu}_A^h = \sin^2 y$, respectively. Then μ is a *Teichmüller differential* on A if it is a constant multiple of μ_A^t and is a *harmonic differential* on A if it is a constant multiple of μ_A^h.

Lemma A.3 *Let μ be a Beltrami differential on Y such that on an annulus A, $\mu = c\mu_A^t$ is a Teichmüller differential. Assume that ν is the Beltrami differential with $\nu = 2c\mu_A^h$ on A and $\nu = \mu$ on the complement of A. Then $\mu - \nu$ is an infinitesimally trivial Beltrami differential.*

Proof We need to show that for any holomorphic quadratic differential $\phi \in Q(Y)$ the pairing of ϕ with $\mu - \nu$ is zero. The difference $\mu - \nu$ is supported on A so our computation will be on fundamental domain in S for the action $z \mapsto z + \pi/m(A)$. The restriction of ϕ to A lifts to a holomorphic quadratic differential $g(z)\,dz^2$ on S, where g is a periodic holomorphic function. That is,

$$g\left(z + \frac{\pi}{m(A)}\right) = g(z).$$

Let

$$b(y) = \int_0^{\pi/m(A)} g(x+iy)\,dx.$$

If Q is a rectangle whose top and bottom sides are horizontal segments from $x = 0$ to $x = \pi/m(A)$ at heights $y_0 < y_1$ then

$$\int_{\partial Q} g(z)\,dz = b(y_0) - b(y_1)$$

since the periodicity of $g(z)$ implies that the line integrals over the vertical sides cancel. As $g(z)$ is holomorphic the line integral around ∂Q is zero and therefore $b(y_0) = b(y_1)$, which implies that $b(y) \equiv b$ is a constant function.

Using this we now compute the pairing:

$$\begin{aligned}\int_Y (\mu - \nu)\phi = \int_A (\mu - \nu)\phi &= \int_0^{\pi} \int_0^{\pi/m(A)} c(1 - 2\sin^2 y) g(x+iy)\,dx\,dy \\ &= \int_0^{\pi} cb(1 - 2\sin^2 y)\,dy = 0.\end{aligned}$$

□

In practice it is easier to construct deformations where the tangent vectors are infinitesimal Teichmüller differentials on annuli. We can use the previous lemma to bound the Weil–Petersson norm of these deformations.

Lemma A.4 *Let A_i be a collection of disjoint annuli on Y with finite moduli m_i. If*

$$\mu = \sum_i c_i \mu^t_{A_i}$$

is a Beltrami differential on Y, then

$$\|[\mu]\|_2^2 \le 2\pi^2 \sum \frac{|c_i|^2}{m_i}.$$

Proof By Lemma A.3, the Beltrami differential μ is equivalent to

$$\nu = 2\sum_i c_i \mu^h_{A_i},$$

so

$$\|[\mu]\|_2^2 = \|[\nu]\|_2^2 \le \int_Y \|\nu\|^2 \, da_Y,$$

where da_Y is the area form for the hyperbolic metric on Y. By the Schwarz lemma if da_i is the area form for the complete hyperbolic metric on A_i then $da_Y < da_i$. On the strip S the area form da_i lifts to $(1/\sin^2 y)\, dx\, dy$ so

$$\begin{aligned}\int_Y \|\nu\|^2 \, da_Y &\le 4\sum_i \int_{A_i} |c_i|^2 \|\mu^h_{A_i}\|^2 \, da_i \\ &= 4\sum_i |c_i|^2 \int_0^{\pi} \int_0^{\pi/m(A_i)} \frac{(\sin^2 y)^2}{\sin^2 y} \, dx\, dy \\ &= 4\sum \frac{|c_i|^2 \pi^2}{2m_i}.\end{aligned}$$

□

We can now describe the strategy of the proof of Proposition A.1. Let $Z \subset Y$ be the complement of the geodesic representatives of τ in Y. Then Z will lift to $\check{Z}$ and conformally embed in both Y and $\widehat{Y}$. We will construct a family of quasiconformal deformations of $\widehat{Y}$ to itself, where the tangent vectors of these deformations will be Teichmüller differentials on a collection of annuli that lie in $Z \subset \widehat{Y}$. As Z is also a subsurface of Y this will define a family of quasiconformal deformations of Y, but here the surface will change along the deformation. This will define a path in $\mathrm{Teich}(S)$. We will use Lemma A.2 to see that this path converges to $\widehat{Y}$ and Lemma A.4 to bound above the Weil–Petersson length of the path.

The cusp deformation Every cusp $\mathfrak{C}$ of a hyperbolic Riemann surface can be parametrized as the quotient of the horodisk

$$\mathfrak{H} = \{z \in \mathbb{C} \mid \operatorname{Im} z \geq 1\}$$

by the translation

$$z \mapsto z + 2.$$

If we let

$$\mathfrak{H}(m) = \{z \in \mathbb{C} \mid 1 \leq \operatorname{Im} z \leq 2m + 1\},$$

then the quotient $\mathfrak{C}(m)$ of $\mathfrak{H}(m)$ is an annulus of modulus m. Define maps

$$f_t^m : \mathfrak{C} \to \mathfrak{C}$$

such that f_t^m is

- constant in the x–variable,
- an affine map from $\mathfrak{C}(m)$ to $\mathfrak{C}(e^t m)$,
- is conformal in the complement of $\mathfrak{C}(m)$.

At time t the infinitesimal Beltrami differential ν_t for this path will be supported on the annulus $\mathfrak{C}(e^t m)$ and using the fact that

$$f_s^{e^t m} \circ f_t^m = f_{s+t}^m,$$

we see that the lift of ν_t to $\mathfrak{H}$ is supported on $\mathfrak{H}(e^t m)$ with $\widetilde{\nu}_t = -\frac{1}{2}$. In particular, ν_t is a Teichmüller differential on $\mathfrak{C}(e^t m)$.

The deformation of $\widehat{Y}$ and Y Each curve in τ is a node of $\widehat{Y}$ and there are two associated cusps in $\widehat{Y}$. If τ has k curves we label the two cusps associated to the i^{th} node by $\mathfrak{C}_i^{\pm}$ and assume that the modulus m_i has been chosen such that the annuli $\mathfrak{C}_i^{\pm}(m_i)$ lie in Z.

With this choice of moduli we define a family of maps

$$f_t : \widehat{Y} \to \widehat{Y}$$

by setting f_t to be the map $f_t^{m_i}$ on the cusps $\mathfrak{C}_i^{\pm}$ and to be the identity on the complement of the cusps. (There may be cusps of $\widehat{Y}$ that don't correspond to nodes in τ. The map is the identity here.) The Beltrami differentials μ_t for this family of maps are supported on the annuli $\mathfrak{C}_i^{\pm}(m_i)$. As these lie in Z, the μ_t are also a family of Beltrami differentials on Y so we have two one-parameter families of surfaces Z_t and Y_t with Z_t conformally embedding in Y_t. The Z_t also conformally embed in $\widehat{Y}$.

Proof of Proposition A.1 Let β_i be the i^{th} curve of τ and let $\beta_i^{\pm}$ be the two curves that are homotopically distinct in $S\backslash\tau$ but are both homotopic in S to β_i. Let $Z_t^{\beta_i^{\pm}}$ be the annular cover of the component of Z_t containing $\beta_i^{\pm}$. Then

$$m(Z_t^{\beta_i^{\pm}}) \geq e^t m_i$$

and therefore

$$\ell_{\partial Z_t}(Z_t) \to 0 \quad \text{as } t \to \infty.$$

By Lemma A.2 for all nonperipheral simple closed curves γ in R we have

$$\lim_{t\to\infty} \ell_\gamma(Z_t) = \lim_{t\to\infty} \ell_\gamma(Y_t), \qquad \lim_{t\to\infty} \ell_\gamma(Z_i) = \lim_{t\to\infty} \ell_\gamma(\widehat{Y}) = \ell_\gamma(\widehat{Y}).$$

It follows that

$$\lim_{t\to\infty} \ell_\gamma(Y_t) = \ell_\gamma(\widehat{Y}),$$

so $Y_t \to \widehat{Y}$ in $\overline{\text{Teich}(S)}$.

The tangent vector of the path are Teichmüller differentials on $2k$ disjoint annuli with coefficients $-\frac{1}{2}$. At time t, two of these annuli have modulus $e^t m_i$, so integrating the estimate from Lemma A.4 we have

$$d_{\text{WP}}(Y, \widehat{Y}) \leq \int_0^\infty \sqrt{\pi^2 \sum_i \frac{1}{m_i e^t}} = 2\pi \sqrt{\sum \frac{1}{m_i}}.$$

To finish the proof we need to bound the m_i from below. As $\check{Z}$ is a cover of Y, $\ell_{\beta_i^{\pm}}(\check{Z}) = \ell_{\beta_i}(Y)$. By the Schwarz lemma, the geodesic representative of $\beta_i^{\pm}$ in $\check{Z}$ will lie in the $\ell_{\beta_i^{\pm}}(\check{Z})/2$–thin part of the associated cusps $\mathfrak{C}^{\pm}$ of $\widehat{Y}$. If $p \in \mathfrak{C}$ is a point in our standard model of a cusp with pre-image $z = x + iy \in \mathfrak{H}$ then injectivity radius satisfies the formula

$$\sinh(\text{inj}(p)) = \frac{1}{y}.$$

Note that while z is not uniquely determined, the y–coordinate is. This implies that $\check{Z}$ will contain the annuli $\mathfrak{C}(m_i)$ where

$$m_i = \frac{1}{2}\left(\frac{1}{\sinh(\ell_{\beta_i}(Y)/2)} - 1\right).$$

With our assumption that $\ell_{\beta_i^{\pm}}(\check{Z}) = \ell_{\beta_i}(Y) \leq \ell_0$ we have

$$\sinh(\ell_{\beta_i}(Y)/2) \leq \frac{\sinh(\ell_0/2)}{\ell_0/2} \cdot \frac{\ell_{\beta_i}(Y)}{2}$$

and therefore

$$\begin{aligned} m_i &\geq \frac{\ell_0}{2\sinh(\ell_0/2)\ell_{\beta_i}(Y)} - \frac{1}{2} = \frac{\ell_0 - \sinh(\ell_0/2)\ell_{\beta_i}(Y)}{2\sinh(\ell_0/2)\ell_{\beta_i}(Y)} \\ &\geq \frac{\ell_0 - \sinh(\ell_0/2)\ell_0}{2\sinh(\ell_0/2)\ell_{\beta_i}(Y)} = \frac{\ell_0(1-\sinh(\ell_0/2))}{2\sinh(\ell_0/2)\ell_{\beta_i}(Y)}. \end{aligned}$$

It follows that

$$d_{\mathrm{WP}}(Y,\hat{Y}) \leq 2\pi\sqrt{\sum_i \frac{2\sinh(\ell_0/2)\ell_{\beta_i}(Y)}{\ell_0(1-\sinh(\ell_0/2))}} = 2\pi\sqrt{\frac{2\sinh(\ell_0/2)}{\ell_0(1-\sinh(\ell_0/2))}}\sqrt{\ell_\tau(Y)}. \quad \square$$

References

[1] **T Aougab**, **S J Taylor**, **R C H Webb**, *Effective Masur–Minsky distance formulas and applications to hyperbolic* 3*–manifolds*, preprint (2017) Available at `https //gauss.math.yale.edu/s̃t654/main.pdf`

[2] **D E Barrett**, **J Diller**, *Contraction properties of the Poincaré series operator*, Michigan Math. J. 43 (1996) 519–538 MR Zbl

[3] **F Bonahon**, **J-P Otal**, *Laminations measurées de plissage des variétés hyperboliques de dimension* 3, Ann. of Math. 160 (2004) 1013–1055 MR Zbl

[4] **M Bridgeman**, **J Brock**, **K Bromberg**, *Schwarzian derivatives, projective structures, and the Weil–Petersson gradient flow for renormalized volume*, Duke Math. J. 168 (2019) 867–896 MR Zbl

[5] **M Bridgeman**, **K Bromberg**, *L^2–bounds for drilling short geodesics in convex cocompact hyperbolic* 3*–manifolds*, preprint (2021) arXiv 2112.02724

[6] **M Bridgeman**, **K Bromberg**, *A bound on the L^2–norm of a projective structure by the length of the bending lamination*, preprint (2022) arXiv 2201.08926

[7] **M Bridgeman**, **K Bromberg**, *Strata separation for the Weil–Petersson completion and gradient estimates for length functions*, J. Topol. Anal. (online publication January 2022)

[8] **M Bridgeman**, **Y Wu**, *Uniform bounds on harmonic Beltrami differentials and Weil–Petersson curvatures*, J. Reine Angew. Math. 770 (2021) 159–181 MR Zbl

[9] **J F Brock**, *The Weil–Petersson metric and volumes of* 3*–dimensional hyperbolic convex cores*, J. Amer. Math. Soc. 16 (2003) 495–535 MR Zbl

[10] **J F Brock**, *Weil–Petersson translation distance and volumes of mapping tori*, Comm. Anal. Geom. 11 (2003) 987–999 MR Zbl

[11] **J F Brock**, **K W Bromberg**, *Inflexibility, Weil–Peterson distance, and volumes of fibered* 3*–manifolds*, Math. Res. Lett. 23 (2016) 649–674 MR Zbl

[12] **J F Brock**, **R D Canary**, **Y N Minsky**, *The classification of Kleinian surface groups, II: The ending lamination conjecture*, Ann. of Math. 176 (2012) 1–149 MR Zbl

[13] **J Brock**, **H Masur**, *Coarse and synthetic Weil–Petersson geometry: quasi-flats, geodesics and relative hyperbolicity*, Geom. Topol. 12 (2008) 2453–2495 MR Zbl

[14] **J Brock**, **Y Minsky**, **H Namazi**, **J Souto**, *Bounded combinatorics and uniform models for hyperbolic* 3*–manifolds*, J. Topol. 9 (2016) 451–501 MR Zbl

[15] **P Buser**, *The collar theorem and examples*, Manuscripta Math. 25 (1978) 349–357 MR Zbl

[16] **Y-E Choi**, **C Series**, *Lengths are coordinates for convex structures*, J. Differential Geom. 73 (2006) 75–117 MR Zbl

[17] **C R Graham**, **E Witten**, *Conformal anomaly of submanifold observables in AdS/CFT correspondence*, Nuclear Phys. B 546 (1999) 52–64 MR Zbl

[18] **L Keen**, *Collars on Riemann surfaces*, from "Discontinuous groups and Riemann surfaces" (L Greenberg, editor), Ann. of Math. Stud. 79, Princeton Univ. Press (1974) 263–268 MR Zbl

[19] **S Kojima**, **G McShane**, *Normalized entropy versus volume for pseudo-Anosovs*, Geom. Topol. 22 (2018) 2403–2426 MR Zbl

[20] **I Kra**, *Deformation spaces*, from "A crash course on Kleinian groups" (L Bers, I Kra, editors), Lecture Notes in Math. 400, Springer (1974) 48–70 MR Zbl

[21] **K Krasnov**, **J-M Schlenker**, *On the renormalized volume of hyperbolic* 3*–manifolds*, Comm. Math. Phys. 279 (2008) 637–668 MR Zbl

[22] **K Krasnov**, **J-M Schlenker**, *The Weil–Petersson metric and the renormalized volume of hyperbolic* 3*–manifolds*, from "Handbook of Teichmüller theory, III" (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 17, Eur. Math. Soc., Zürich (2012) 779–819 MR Zbl

[23] **O Lehto**, *Univalent functions and Teichmüller spaces*, Graduate Texts in Math. 109, Springer (1987) MR Zbl

[24] **M Linch**, *A comparison of metrics on Teichmüller space*, Proc. Amer. Math. Soc. 43 (1974) 349–352 MR Zbl

[25] **H Masur**, *Extension of the Weil–Petersson metric to the boundary of Teichmuller space*, Duke Math. J. 43 (1976) 623–635 MR Zbl

[26] **C McMullen**, *Iteration on Teichmüller space*, Invent. Math. 99 (1990) 425–454 MR Zbl

[27] **J W Morgan**, *On Thurston's uniformization theorem for three-dimensional manifolds*, from "The Smith conjecture" (J W Morgan, H Bass, editors), Pure Appl. Math. 112, Academic (1984) 37–125 MR Zbl

[28] **F V Pallete**, *Continuity of the renormalized volume under geometric limits*, preprint (2016) arXiv 1605.07986

[29] **J-M Schlenker**, *The renormalized volume and the volume of the convex core of quasi-fuchsian manifolds*, Math. Res. Lett. 20 (2013) 773–786 MR Zbl

[30] **P A Storm**, *Minimal volume Alexandrov spaces*, J. Differential Geom. 61 (2002) 195–225 MR Zbl

[31] **P A Storm**, *Hyperbolic convex cores and simplicial volume*, Duke Math. J. 140 (2007) 281–319 MR Zbl

[32] **L A Takhtajan**, **L-P Teo**, *Liouville action and Weil–Petersson metric on deformation spaces, global Kleinian reciprocity and holography*, Comm. Math. Phys. 239 (2003) 183–240 MR Zbl

[33] **W P Thurston**, *Hyperbolic structures on* 3*–manifolds, III: Deformations of* 3*–manifolds with incompressible boundary*, preprint (1998) arXiv math/9801058

[34] **S A Wolpert**, *Geometry of the Weil–Petersson completion of Teichmüller space*, from "Surveys in differential geometry, VIII" (S-T Yau, editor), Surv. Differ. Geom. 8, Int., Somerville, MA (2003) 357–393 MR Zbl

[35] **P G Zograf**, **L A Takhtajan**, *On the uniformization of Riemann surfaces and on the Weil–Petersson metric on the Teichmüller and Schottky spaces*, Mat. Sb. 132(174) (1987) 304–321, 444 MR Zbl In Russian; translated in Math. USSR-Sb. 60 (1988) 297–313

Department of Mathematics, Boston College
Chestnut Hill, MA, United States

Department of Mathematics, Yale University
New Haven, CT, United States

Department of Mathematics, University of Utah
Salt Lake City, UT, United States

bridgem@bc.edu, jeffrey.brock@yale.edu, bromberg@math.utah.edu

https://sites.google.com/bc.edu/martin-bridgeman,
http://www.jeffbrock.net, http://math.utah.edu/~bromberg

Proposed: Benson Farb — Received: 1 February 2021
Seconded: Tobias H Colding, David Gabai — Revised: 21 January 2022

msp

Geometry & Topology **27**:8 (2023) 3229–3302
DOI: 10.2140/gt.2023.27.3229
Published: 9 November 2023

Weighted K–stability and coercivity with applications to extremal Kähler and Sasaki metrics

Vestislav Apostolov
Simon Jubert
Abdellah Lahdili

We show that a compact weighted extremal Kähler manifold, as defined by the third author (2019), has coercive weighted Mabuchi energy with respect to a maximal complex torus $\mathbb{T}^{\mathbb{C}}$ in the reduced group of complex automorphisms. This provides a vast extension and a unification of a number of results concerning Kähler metrics satisfying special curvature conditions, including Kähler metrics with constant scalar curvature, extremal Kähler metrics, Kähler–Ricci solitons, and their weighted extensions. Our result implies the strict positivity of the weighted Donaldson–Futaki invariant of any nonproduct $\mathbb{T}^{\mathbb{C}}$–equivariant smooth Kähler test configuration with reduced central fibre, a property known as $\mathbb{T}^{\mathbb{C}}$–equivariant weighted K–polystability on such test configurations. It also yields the $\mathbb{T}^{\mathbb{C}}$–uniform weighted K–stability on the class of smooth $\mathbb{T}^{\mathbb{C}}$–equivariant polarized test configurations with reduced central fibre. For a class of fibrations constructed from principal torus bundles over a product of Hodge cscK manifolds, we use our results in conjunction with results of Chen and Cheng (2021), He (2019) and Han and Li (2022) in order to characterize the existence of extremal Kähler metrics and Calabi–Yau cones associated to the total space, in terms of the coercivity of the weighted Mabuchi energy of the fibre. This yields a new existence result for Sasaki–Einstein metrics on certain Fano toric fibrations, extending the results of Futaki, Ono and Wang (2009) in the toric Fano case, and of Mabuchi and Nakagawa (2013) in the case of Fano $\mathbb{P}^1$–bundles.

32Q20, 53C25, 53C55, 58J60; 14J45, 32J27

Introduction

We are concerned with the existence and obstruction theory of a class of special Kähler metrics, called *weighted constant scalar curvature metrics*, which were introduced by the third author in [54; 55], giving a vast extension of the notion of Kähler metrics of constant scalar curvature (cscK for short), and providing the unification of a number of related notions of Kähler metrics satisfying special curvature conditions.

0.1 The weighted cscK problem

Let X be a smooth compact complex m–dimensional manifold with a given de Rham cohomology class $\alpha \in H^{1,1}(X,\mathbb{R})$ of Kähler metrics, and let $\mathbb{T} \subset \mathrm{Aut}_r(X)$ denote a fixed compact torus in the *reduced* group $\mathrm{Aut}_r(X)$ of automorphisms of X, ie the connected subgroup of automorphisms of X generated by the Lie algebra of real holomorphic vector fields with zeros; see eg Gauduchon [41]. It is well known that $\mathbb{T}$ acts in a hamiltonian way with respect to any $\mathbb{T}$–invariant Kähler metric $\omega \in \alpha$, and the corresponding momentum map μ_ω sends X onto a compact convex polytope $\Delta \subset \mathfrak{t}^*$ in the dual vector space $\mathfrak{t}^*$ of the Lie algebra $\mathfrak{t}$ of $\mathbb{T}$; see Atiyah [9] and Guillemin and Sternberg [44]. Furthermore, up to translations, Δ is independent of the choice of $\omega \in \alpha$. We shall further fix Δ, giving rise to a *normalization* of the corresponding momentum maps $\{\mu_\omega \mid \omega \in \alpha\}$.

Following [55], let $v(\mu) > 0$ and $w(\mu)$ be smooth functions defined on Δ. One can then consider the following condition for $\mathbb{T}$–invariant Kähler metrics ω in α (and fixed polytope Δ), called (v,w)*–cscK metrics*:

$$\mathrm{Scal}_v(\omega) = w(\mu_\omega). \tag{1}$$

Here the so-called v*–scalar curvature* of ω is defined by

$$\mathrm{Scal}_v(\omega) := v(\mu_\omega)\,\mathrm{Scal}(\omega) + 2\Delta_\omega v(\mu_\omega) + \langle g_\omega, \mu_\omega^*(\mathrm{Hess}(v))\rangle, \tag{2}$$

with $\mathrm{Scal}(\omega)$ being the usual scalar curvature of the riemannian metric g_ω associated to ω, Δ_ω the Laplace operator of g_ω, and where the contraction $\langle\cdot,\cdot\rangle$ is taken between the smooth $\mathfrak{t}^*\otimes\mathfrak{t}^*$–valued function g_ω on X (the restriction of the riemannian metric g_ω to $\mathfrak{t} \subset C^\infty(X,TX)$) and the smooth $\mathfrak{t}\otimes\mathfrak{t}$–valued function ${\mu_\omega}^*(\mathrm{Hess}(v))$ on X (given by the pullback by μ_ω of $\mathrm{Hess}(v) \in C^\infty(\Delta, \mathfrak{t}\otimes\mathfrak{t})$). The relevance of (1) to various geometric conditions is discussed in detail in [55], but we mention below a few special cases which partly motivate our study:

- $v = 1$ and w is a constant: this is the familiar cscK problem.

- $v = 1$ and $w = \ell$ with ℓ an affine-linear function on $\mathfrak{t}^*$: (1) then describes an extremal Kähler metric in the sense of Calabi [21].

- $v = e^\ell$ and $w = 2(\ell + a)e^\ell$, where ℓ is an affine-linear function on $\mathfrak{t}^*$ and a is a constant corresponding to the so-called μ–cscK (see Inoue [50]), extending the notion of Kähler–Ricci solitons (see Tian and Zhu [71]) defined when X is Fano and $\alpha = 2\pi c_1(X)$.

- $v = \ell^{-m-1}$, $w = a\ell^{-m-2}$ and $\alpha = c_1(L)$, where ℓ is a positive affine-linear function on Δ, m is the complex dimension of X, a is a constant, and L is a polarization of X: (1) then describes a scalar flat cone Kähler metric on the affine cone $(L^{-1})^{\times}$ polarized by the lift of $\xi = d\ell$ to L^{-1} via ℓ; see Apostolov, Calderbank and Legendre [2; 7].

In general, the problem of finding a $\mathbb{T}$–invariant Kähler metric $\omega \in \alpha$ solving (1) is obstructed in a similar way that the cscK problem is obstructed by the vanishing of the Futaki invariant: for any $\mathbb{T}$–invariant Kähler metric $\omega \in \alpha$ and any affine-linear function ℓ on $\mathfrak{t}^*$, one must have

$$\mathrm{Fut}_{v,w}(\ell) := \int_X (\mathrm{Scal}_v(\omega) - w(\mu_\omega))\ell(\mu_\omega)\omega^m = 0, \tag{3}$$

should a solution to (1) exist. In [55], an unobstructed modification of (1) is proposed, extending Calabi's notion [21] of extremal Kähler metrics. To this end, suppose that $v, w_0 > 0$ are positive smooth functions on Δ. One can then find a unique affine-linear function $\ell^{\mathrm{ext}}_{v,w_0}(\mu)$ on $\mathfrak{t}^*$, called *the extremal function*, such that (3) holds for the weights $(v, w) = (v, \ell^{\mathrm{ext}}_{v,w_0} w_0)$. In this case, a solution of the (v, w)–cscK problem (1) is referred to as a (v, w_0)*–extremal Kähler metric*. We emphasize that (v, w_0)–extremal Kähler metrics are (v, w)–cscK metrics with a special property of the weight function w, namely, $w = \ell w_0$ with $w_0 > 0$ on Δ and ℓ affine-linear. In particular, (v, w)–cscK metrics with $w \neq 0$ on Δ are (v, w)–extremal with $\ell^{\mathrm{ext}}_{v,w} = \mathrm{sign}(w|_\Delta)$ and $(v, 0)$–cscK metrics are (v, w)–extremal with $\ell^{\mathrm{ext}}_{v,w} = 0$ for any $w > 0$. It follows that all the above listed special cases are examples of (v, w)–extremal Kähler metrics, and thus the setup of (v, w)–extremal Kähler metrics allows one to study all these cases together.

0.2 Relation to v–solitons

Motivated by works of T Mabuchi [58; 59] and subsequent work by Berman and Nyström [15], Y Han and C Li [45] have recently introduced and studied the general notion of a *weighted* v*–soliton* on a smooth Fano variety X, as follows. In the setup explained above, we let $\alpha = 2\pi c_1(X)$ and consider the natural action of $\mathbb{T}$ on K_X^{-1}, which fixes the momentum polytope Δ of $(X, \alpha, \mathbb{T})$ and normalizes the momentum map μ_ω for any $\mathbb{T}$–invariant Kähler metric $\omega \in \alpha$. For a (smooth) positive weight function $v(\mu)$ on Δ, one defines a v–soliton as a $\mathbb{T}$–invariant Kähler metric $\omega \in \alpha$, such that

$$\rho_\omega - \omega = \tfrac{1}{2} dd^c \log v(\mu_\omega), \tag{4}$$

where ρ_ω denotes the Ricci form of ω. Notice that when $v(\mu) = e^{\langle \mu, \xi \rangle}$ for some $\xi \in \mathfrak{t}$, one gets the well-studied class of *Kähler–Ricci solitons* [71] whereas the case when

$v(\mu)$ is a positive affine-linear function on Δ corresponds to the *Mabuchi solitons* studied in [58; 59]. As we shall see below, other choices for v are also geometrically meaningful. We make the following useful observation:

Proposition 1 *Let X be a smooth Fano manifold and $\mathbb{T} \subset \mathrm{Aut}(X)$ a compact torus. A $\mathbb{T}$–invariant Kähler metric $\omega \in 2\pi c_1(X)$ is a v–soliton if and only if ω is (v,w)–cscK with $w(\mu) := 2(m + \langle d \log v, \mu\rangle)v(\mu)$.*

We use the above result in order to make connection with the recent paper [45] (where the authors obtain a complete Yau–Tian–Donaldson type correspondence for the existence of v–Ricci solitons), which will play an important role in our present study of (v,w)–cscK metrics.

We also notice that v–solitons can be viewed as $(\bar v, \bar w)$–cscK metrics for *different* choices of weights. This is for instance the case when $v(\mu) = \ell(\mu)^{-(m+2)}$, where $\ell(\mu) = \langle \xi, \mu\rangle + a$ is positive affine-linear on Δ. Whereas Proposition 1 identifies the v–soliton as a (v,w)–cscK metric with

$$v = \ell^{-(m+2)} \quad \text{and} \quad w = 2\ell^{-(m+3)}(-2\ell + (m+2)a),$$

we also observe:

Proposition 2 *Let $(X, \mathbb{T})$ be a smooth Fano variety and $\ell(\mu) = (\langle \xi, \mu\rangle + a)$ a positive affine-linear function on its canonical polytope Δ. A $\mathbb{T}$–invariant Kähler metric $\omega \in 2\pi c_1(X)$ is an $\ell^{-(m+2)}$–soliton if and only if the lift $\hat\xi$ of $\xi = d\ell$ to K_X via ℓ is the Reeb vector field of a Sasaki–Einstein structure defined on the unit circle bundle $N \subset K_X$ with respect to the hermitian metric on K_X with curvature $-\omega$. This condition is also equivalent to ω being an $(\ell^{-m-1}, 2ma\ell^{-m-2})$–cscK metric.*

0.3 Main results

Similarly to the usual cscK case, it is shown by Lahdili [55] that the solutions of (1) can be characterized as minimizers of a functional $\boldsymbol{M}_{v,w}$ defined on the space of $\mathbb{T}$–invariant Kähler metrics in α, extending the *Mabuchi energy* to the weighted setting (see Section 1 below for the precise definition). After the deep works of Berman, Darvas and Lu [14] and Chen and Cheng [23], it is now well understood that the coercivity of the Mabuchi energy is equivalent to the existence of a cscK metric in a given cohomology class. Noting that, by the results in [55], any (v,w)–extremal metric is invariant under a *maximal* compact torus in $\mathrm{Aut}_r(X)$, our first main result is an extension of one direction of the correspondence in the cscK case to the weighted setting.

Theorem 1 *Suppose $\mathbb{T} \subset \mathrm{Aut}_r(X)$ is a maximal torus in the reduced group of automorphisms of X, and $\omega_0 \in \alpha$ a $\mathbb{T}$–invariant (v, w_0)–extremal Kähler metric. Then the weighted Mabuchi energy $\boldsymbol{M}_{v,w}$ (with $w = \ell^{\mathrm{ext}}_{v,w_0} w_0$) is coercive relative to the complex torus $\mathbb{T}^{\mathbb{C}}$ in the sense of Darvas and Rubinstein* [29], *ie there exist positive real constants λ and δ such that for any $\mathbb{T}$–invariant Kähler metric $\omega \in \alpha$,*

$$\boldsymbol{M}_{v,w}(\omega) \geq \lambda \inf_{\sigma \in \mathbb{T}^{\mathbb{C}}} \boldsymbol{J}(\sigma^*\omega) - \delta,$$

where $\boldsymbol{J}$ denotes the Aubin functional on the space of Kähler metrics; *see Definition 3.1.*

Our proof of Theorem 1 adapts to the case when the torus $\mathbb{T} \subset \mathrm{Aut}_r(X)$ is not necessarily maximal. Instead of $\mathbb{T}^{\mathbb{C}}$ one takes the infimum of $\boldsymbol{J}(\sigma^*\omega)$ over $\widehat{\mathbb{G}} := \mathrm{Aut}_r^{\mathbb{T}}(X)$, the connected component of the identity of the centralizer of $\mathbb{T}$ in $\mathrm{Aut}_r(X)$ (which by [55] is a reductive group if X admits a (v, w_0)–extremal $\mathbb{T}$–invariant Kähler metric; see Remark 7.7 for more details). Furthermore, we can also consider *any* reductive connected subgroup group $\mathbb{G} = \mathbb{K}^{\mathbb{C}} \subset \widehat{\mathbb{G}}$ with a compact form $\mathbb{K}$ containing $\mathbb{T}$, and restrict $\boldsymbol{M}_{v,w}$ to the space of $\mathbb{K}$–invariant Kähler metrics in α as in Han and Li [45].[1]

As noticed by Berman, Darvas and Lu [14] (in the polarized case) and by Sjöström Dyrefelt [66] (in the more general Kähler case), the coercivity of the Mabuchi energy yields a sharp estimate of the sign of the Donaldson–Futaki invariant of a $\mathbb{T}$–equivariant test configuration. In our weighted setting, we consider $\mathbb{T}$–equivariant (compactified) Kähler test configurations $(\mathscr{X}, \mathscr{A})$ associated to $(X, \alpha, \mathbb{T})$, which have smooth total space. To any such test configuration one can associate a *weighted Donaldson–Futaki* invariant by the formula (see [55])

$$\mathscr{F}_{v,w}(\mathscr{X}, \mathscr{A}) := -\int_{\mathscr{X}} (\mathrm{Scal}_v(\Omega) - w(\mu_\Omega))\Omega^{[m+1]} + (8\pi) \int_X v(\mu_\omega)\omega^{[m]},$$

where $\Omega \in \mathscr{A}$ and $\omega \in \alpha$ are $\mathbb{T}$–invariant Kähler forms on $\mathscr{X}$ and X, respectively, with respective Δ–normalized momentum maps μ_Ω and μ_ω, and $\mathrm{Scal}_v(\Omega)$ is the v–scalar curvature of Ω defined by (2). In the above formula, for any 2–form ψ we use the convention $\psi^{[k]} := \psi^k/k!$. Thus, $\mathscr{F}_{v,w}(\mathscr{X}, \mathscr{A})$ extends to the weighted setting the expression of the Donaldson–Futaki invariant of $(\mathscr{X}, \mathscr{A})$ in terms of intersection numbers (see Odaka [62] and Wang [72]).

Corollary 1 *Under the hypotheses of Theorem 1, for any $\mathbb{T}$–equivariant smooth Kähler test configuration $(\mathscr{X}, \mathscr{A})$ of $(X, \alpha, \mathbb{T})$ which has a reduced central fibre,*

$$\mathscr{F}_{v,w}(\mathscr{X}, \mathscr{A}) \geq 0,$$

[1] We are grateful to Chi Li for pointing this out to us.

with equality if and only if $(\mathscr{X},\mathscr{A})$ *is a product test configuration. Furthermore, if* $\alpha = 2\pi c_1(L)$ *corresponds to a polarization* L *of* X *and* $(\mathscr{X},\mathscr{L},\mathbb{T})$ *is a* $\mathbb{T}$*–equivariant smooth polarized test configuration of* (X, L) *as above,*

$$\mathscr{F}_{v,w}(\mathscr{X},\mathscr{A}) \geq \lambda \boldsymbol{J}^{\mathrm{NA}}_{\mathbb{T}^{\mathbb{C}}}(\mathscr{X},\mathscr{A}),$$

where $\mathscr{A} = 2\pi c_1(\mathscr{L})$, $\lambda > 0$ *is the constant appearing in Theorem 1, and* $\boldsymbol{J}^{\mathrm{NA}}_{\mathbb{T}^{\mathbb{C}}}(\mathscr{X},\mathscr{A})$ *is the* $\mathbb{T}^{\mathbb{C}}$*–relative non-Archimedean* $\boldsymbol{J}$*–functional of the test configuration introduced in Hisamoto* [49] *and Li* [57]; *see* (20).

Corollary 1 improves the ($\mathbb{T}$–equivariant) (v, w)–K–semistability established in Lahdili [56, Theorem 2] to ($\mathbb{T}$–equivariant) (v, w)–K–polystability on the test configurations as above, and, in the projective case, further to $\mathbb{T}^{\mathbb{C}}$–uniform (v, w)–K–stability in the sense of [49; 57]. As we already mentioned, the fist part of Corollary 1 was proved by Berman, Darvas and Lu [14], Sjöström Dyrefelt [66] and Stoppa [67] in the cscK case ($v = 1$ and w is a constant), and by Dervan [33], [66] and Stoppa and Székelyhidi [68] in the unweighted extremal case ($v = 1 = w_0$). We however notice that in the extremal case our proof uses directly the coercivity of the *relative* Mabuchi energy (which follows from Theorem 1) whereas the proofs in [33; 68] and [66] are based on the Arezzo–Pacard existence results of extremal metrics on blow-ups (see Arezzo, Pacard and Singer [8]), and on the coercivity of the unweighted Mabuchi energy $\boldsymbol{M}_{1,c}$ established in [14; 66], respectively. The $\mathbb{T}^{\mathbb{C}}$–uniform (v, w)–K–stability statement in the second part of Corollary 1 is established in the cscK case in [49; 57], and in the case of a v–soliton by Han and Li [45]. Our proof of Corollary 1 in the general weighted case follows easily from Theorem 1 by the established techniques in the cscK case; see Section 4.

Another notable special case where our results apply is when $\alpha = c_1(L)$ for an ample line bundle L over X, and $v = \ell^{-m-1}$ and $w_0 = \ell^{-m-3}$ for a positive affine-linear function on Δ. It is observed by Apostolov and Calderbank [2] that in this case a (v, w_0)–extremal Kähler metric in α describes an *extremal Sasaki* metric on the total space N of the unit circle bundle in L^{-1} with respect to the hermitian metric with curvature $-\omega$, and Reeb vector field corresponding to the lift of $d\ell$ to L^{-1} via ℓ. In this special case, the first part of Corollary 1 above was obtained by Apostolov, Calderbank and Legendre [7] for *polarized* test configurations (see Theorem 1, Conjecture 5.8 and Remark 5.9 in [7]), by using the results in He and Li [48] which establish an analogue of Theorem 1 in the Sasaki case. Thus, our proofs of Theorem 1 and Corollary 1 allow one to recast and further generalize [7, Theorem 1] entirely within the framework of the weighted Kähler geometry of X.

0.4 Method of proof

We now discuss briefly the method of proof of Theorem 1 above. It is an application of the general coercivity principle of Darvas and Rubinstein [29, Theorem 3.4]; see Section 3. This principle is used in the cscK case by Berman, Darvas and Lu [14], and our approach is mainly inspired by these two references. Noting that in the weighted extremal case $\boldsymbol{M}_{v,w}$ is $\mathbb{G}$–invariant and $\mathbb{G} := \mathbb{T}^{\mathbb{C}}$ is reductive, by the results of [29], in order to obtain Theorem 1 one needs to

(i) extend $\boldsymbol{M}_{v,w}$ to the space $\mathcal{E}^1(X, \omega_0)$ of ω_0–relative plurisubharmonic functions of maximal mass and finite energy;

(ii) show that the extension is convex and continuous along weak d_1–geodesics in $\mathcal{E}^1(X, \omega_0)$;

(iii) establish a compactness result for the extension of $\boldsymbol{M}_{v,w}$; and

(iv) show the uniqueness modulo the action of $\mathbb{G}$ (and in particular the regularity) of the weak minimizers of $\boldsymbol{M}_{v,w}$, under the assumption that a (v, w_0)–extremal metric exists.

The steps (i), (ii) and (iii) in the unweighted cscK case are obtained by Berman, Darvas and Lu [13] and follow from the Chen–Tian formula of $\boldsymbol{M}_{1,1}$. The analogous formula for $\boldsymbol{M}_{v,w}$ is obtained by Lahdili [55], but the presence of weights does not allow for a straightforward generalization of the arguments in [13]. Similar difficulty arises in Berman and Nyström [15] in the framework of v–solitons on a Fano variety, where the authors were able to obtain a suitable extension of the *weighted Ding functional* to the space $\mathcal{E}^1(X, \omega_0)$. This functional has milder dependence on the weights than the weighted Mabuchi functional we consider. Indeed, the arguments of [15] yield the existence of a continuous extension to $\mathcal{E}^1(X, \omega_0)$ of one of the three terms in the Chen–Tian decomposition of $\boldsymbol{M}_{v,w}$, which depend on the weight w. Building on [15], Han and Li [45] proposed a new approach to the extension problem in the case of v–solitons, based on an idea going back to Donaldson [36] (see in particular the proof of Proposition 3 in [36]), which amounts to considering suitable fibre bundles Y over a cscK base B and fibre X, and showing that the weighted quantities on X correspond to the restrictions of unweighted quantities on the total space Y. This is the *semisimple principal* $(X, \mathbb{T})$*–fibration construction*, which we review in the next subsection. Going further than [45], we express in general the scalar curvature of a bundle-compatible Kähler metric on Y in terms of the weighted scalar curvature of X, and show that the usual (unweighted) Mabuchi energy on Y restricts to a suitably weighted Mabuchi energy on X. It thus follows that, at least for suitable polynomial weights v, the remaining terms of the Chen–Tian decomposition of $\boldsymbol{M}_{v,w}$ can be extended to $\mathcal{E}^1(X,\omega_0)$ simply by restricting to the fibres the corresponding (unweighted) extension of the Mabuchi energy of Y. The final crucial observation for obtaining the extension for *any* weights is that $\boldsymbol{M}_{v,w}$ depends *linearly* and *continuously* on (v,w), so that one can further use (as in [45]) the Stone–Weierstrass approximation theorem over $C^0(\Delta)$. With this in place, and using the weighted analogue of the uniqueness (see Berman and Berndtsson [11]) achieved by Lahdili [56], we can adapt the arguments from [14].

0.5 Applications to the semisimple principal fibration construction

We briefly review here the semisimple principal bundle construction, which is not only a key tool in our proof of Theorem 1, but also provides a framework for further geometric applications of our results, extending the setting of the generalized Calabi construction in Apostolov, Calderbank, Gauduchon and Tønnesen-Friedman [6].

We denote by $\mathbb{T}$ a compact r–dimensional torus with Lie algebra $\mathfrak{t}$ and lattice $\Lambda\subset\mathfrak{t}$ of generators of S^1–subgroups, ie $\mathbb{T}=\mathfrak{t}/2\pi\Lambda$. Let $B=B_1\times\cdots\times B_k$ be a $2n$–dimensional cscK manifold which is a product of compact cscK Hodge Kähler $2n_a$–manifolds (B_a,ω_{B_a}) for $a=1,\dots,k$. We then consider a principal $\mathbb{T}$–bundle $\pi\colon P\to B$ endowed with a connection 1–form $\theta\in\Omega^1(P,\mathfrak{t})$ with curvature

$$d\theta=\sum_{a=1}^{k}(\pi^*\omega_{B_a})\otimes p_a\quad\text{for } p_a\in\Lambda.$$

For any smooth compact Kähler $2m$–manifold $(X,\omega_X,\mathbb{T})$, endowed with a hamiltonian isometric action of the torus $\mathbb{T}$ as in the setup above, we can construct the *principal* $(X,\mathbb{T})$*–fibration*

$$Y:=(X\times P)/\mathbb{T}\to B,$$

where the $\mathbb{T}$–action on the product is $\sigma(x,p)=(\sigma^{-1}x,\sigma p)$ for $x\in X$, $p\in P$ and $\sigma\in\mathbb{T}$. Using the chosen connection on P, the almost complex structures on X and B lift to define a CR structure on the product $X\times P$, and thus endow Y with the structure of a $2(m+n)$–dimensional smooth complex manifold. Furthermore, Y comes equipped with an induced holomorphic fibration $\pi\colon Y\to B$, with smooth complex fibres X, and induced fibrewise $\mathbb{T}$–action. Fixing constants $c_a\in\mathbb{R}$ such that, for each $a=1,\dots,k$, the affine-linear function $\langle p_a, \mu\rangle + c_a$ on $\mathfrak{t}^*$ is strictly positive on the momentum image Δ of X, one can define a lifted Kähler metric ω_Y on Y which, pulled back to $X \times P$, has the form

$$\omega_Y := \omega_X + \sum_{a=1}^{k} (\langle p_a, \mu_\omega\rangle + c_a)\pi^*\omega_{B_a} + \langle d\mu_\omega \wedge \theta\rangle,$$

where $\langle\cdot,\cdot\rangle$ stands for the natural pairing of $\mathfrak{t}$ and $\mathfrak{t}^*$. Thus $\langle p_a, \mu_\omega\rangle$ is a smooth function and $\langle d\mu_\omega \wedge \theta\rangle$ is a 2–form on $X \times P$. As we show in Section 5, when ω_X varies in a given Kähler class of X, the corresponding Kähler metric ω_Y will vary in a fixed Kähler class on Y. We also notice that when $(X, \omega_X, \mathbb{T})$ is a smooth *toric* Kähler manifold, the setup above reduces to the theory of *semisimple rigid* toric fibrations studied by Apostolov, Calderbank, Gauduchon and Tønnesen-Friedman [3; 4; 6]. Inspired by these results, we show that the scalar curvature of ω_Y can be expressed in terms of the p–weighted scalar curvature of (X, ω_X), where the weight function $p(\mu)$ is a polynomial depending on the fixed data (p_a, c_a, n_a) of the construction. With this observation in mind, we show that (similarly to the case of semisimple rigid toric fibrations recently studied by Jubert [52]) the recent results Chen and Cheng [23] and He [47] can be used to obtain a converse of Theorem 1 in the case of semisimple principal fibrations.

Theorem 2 *Suppose Y is a semisimple principal $(X, \mathbb{T})$–fibration, with a Kähler metric ω_Y induced by a $\mathbb{T}$–invariant Kähler metric ω_X on X. Suppose, moreover, that $\mathbb{T}$ is a maximal torus in the reduced group of automorphisms $\mathrm{Aut}_r(X)$. Then, the following conditions are equivalent*:

(i) *Y admits an extremal Kähler metric in the Kähler class $[\omega_Y]$.*

(ii) *X admits a $\mathbb{T}$–invariant $(p, \tilde{w})$–cscK metric in the Kähler class $[\omega_X]$, with weights*

$$p(\mu) = \prod_{a=1}^{k} (\langle p_a, \mu\rangle + c_a)^{n_a}, \quad \tilde{w}(\mu) = p(\mu)\biggl(-\sum_{a=1}^{k} \frac{\mathrm{Scal}(\omega_{B_a})}{\langle p_a, \mu\rangle + c_a} + \ell^{\mathrm{ext}}(\mu)\biggr),$$

where ℓ^{ext} is an affine-linear function determined by the condition (3).

(iii) *The weighted Mabuchi energy $\boldsymbol{M}^X_{p,\tilde{w}}$ of $(X, [\omega_X], \mathbb{T})$ is coercive with respect to $\mathbb{T}^{\mathbb{C}}$, where p and $\tilde{w}$ are the weights defined in* (ii).

Compared to the general setting of Dervan and Sektnan [35], the semisimple principal $(X, \mathbb{T})$–fibration (trivially) satisfies the condition of optimal symplectic connection. Accordingly, one can conclude by [35] that $(Y, [\omega_Y])$ admits an extremal Kähler metric, provided that (X, ω_X) is cscK, and if we take large enough constants c_a. As a matter of fact, the conclusion also follows under the more general assumption that (X, ω_X) is extremal, by the proof of [4, Theorem 3]. The novelty of Theorem 2 is therefore in the fact that it gives a precise condition (in terms of X) for the existence of an extremal Kähler metric in a *given* Kähler class $[\omega_Y]$, also revealing that $(X, [\omega_X])$ need not be extremal in general. We finally note that in the case of toric fibre, [52] provides a further equivalence with a certain weighted notion of *uniform K–stability* of the corresponding Delzant polytope.

If all the factors (B_a, ω_{B_a}) of the base are positive Kähler–Einstein manifolds, and the fibre $(X, \mathbb{T})$ is a smooth Fano variety, the semisimple principal $(X, \mathbb{T})$–fibration construction can produce a smooth Fano variety Y for suitable choice of the principal $\mathbb{T}$–bundle over B; see Lemma 5.11. In this case, combining Han and Li [45, Theorem 3.5] with our results:

Theorem 3 *Suppose Y is a Fano semisimple principal $(X, \mathbb{T})$–fibration, obtained from the product of positive Kähler–Einstein Hodge manifolds (B_a, ω_{B_a}) and a smooth Fano fibre $(X, \mathbb{T})$ via Lemma 5.11. Suppose also that $\mathbb{T}$ is a maximal torus in the automorphism group* $\mathrm{Aut}(X)$. *Then Y admits a v–soliton in $2\pi c_1(Y)$, provided that the weighted Mabuchi functional $\boldsymbol{M}^X_{pv,\widetilde{w}}$ of $(X, \mathbb{T}, 2\pi c_1(X))$ is coercive with respect to $\mathbb{T}^{\mathbb{C}}$, where p is the weight defined in Theorem 2*(ii) *and*

$$\widetilde{w} = 2pv(m + \langle d\log v, \mu\rangle + \langle d\log p, \mu\rangle).$$

If, furthermore, the fibre $(X, \mathbb{T})$ is a smooth toric Fano variety, then this equation is equivalent to the vanishing of the Futaki invariant (3) *associated to the weights $(pv, \widetilde{w})$ on X. In particular, any Fano semisimple principal $(X, \mathbb{T})$–fibration with smooth toric Fano fibre $(X, \mathbb{T})$ admits a Kähler–Ricci soliton, and the corresponding affine cone $(K_Y)^\times$ admits a Calabi–Yau cone metric, given by a Sasaki–Einstein structure on a unit circle bundle associated to the canonical bundle K_Y.*

The existence of a Kähler–Ricci soliton in the above setting is essentially known even though we didn't find it explicitly stated in the literature. In the toric case (ie when $Y = X$ and B is a point) the result follows by Wang and Zhu [73] (see also Datar and Székelyhidi [30]), and for $\mathbb{P}^1$–bundles by Apostolov, Calderbank, Gauduchon and Tønnesen-Friedman [5], Dancer and Wang [26] and Koiso [53]. In the general case the result can be obtained from Podestà and Spiro [63], which in turn extends [73] to the framework of multiplicity-free manifolds, but the arguments can be also adapted to the case of semisimple principal $(X, \mathbb{T})$–fibrations; see Apostolov, Calderbank, Gauduchon and Tønnesen-Friedman [6, Remark 7] and Donaldson [37]. Our approach, however, builds on the idea of [30]. There are also related existence results for Kähler–Ricci solitons on spherical manifolds; see Delcroix [31] and Delgove [32]. On the other hand, the existence of Sasaki–Einstein metrics seems to be new in the above stated generality. Indeed, in the toric case the claim follows from Futaki, Ono and Wang [40], and there are known existence results (see Boyer and Tønnesen-Friedman [20], Gauntlett, Martelli, Sparks and Waldram [42] and Mabuchi and Nakagawa [60]) on $\mathbb{P}^1$–bundles. We expect our arguments to extend to spherical manifolds too.

0.6 Structure of the paper

In Section 1, we recall the setup of weighted cscK metrics and state the main results we shall need from Lahdili [55; 56]. In Section 2, we recall the notion of v–solitons from Han and Li [45] and Mabuchi [58], and establish the equivalences stated in Propositions 1 and 2. Sections 3 and 4 review and recast in the weighted setting, respectively, the coercivity principle of Darvas and Rubinstein [29] and its application to stability (see Berman, Darvas and Lu [14] and Sjöström Dyrefelt [66]), thus outlining the main steps needed for the proof of Theorem 1 and from it deriving Corollary 1. In Section 5, we introduce the semisimple principal $(X, \mathbb{T})$–fibration construction, and establish the main geometric properties allowing us to extend the results from Apostolov, Calderbank, Gauduchon and Tønnesen-Friedman [6]. In Section 6, we use an idea from [45] in order to define an extension of the weighted Mabuchi energy to the space $\mathcal{E}^1(X, \omega_0)$, and show its convexity and compactness properties. In Section 7, we extend the arguments of [14] to show that weak minimizers of the weighted Mabuchi energy are smooth. Here, we complete the proof of Theorem 1. In Section 8, we detail the proofs of Theorems 2 and 3. In the appendices, we present some technical computational results, detailing the linearization of the scalar and the twisted scalar curvature of a semisimple principal $(X, \mathbb{T})$–fibre and recasting the weighted Futaki invariant (3), which are needed for the proofs of Theorems 2 and 3.

Acknowledgements Apostolov was supported in part by an NSERC discovery grant and a connect talent grant of the Région Pays de la Loire. Jubert was supported by PhD fellowships of the UQAM and the University of Toulouse III – Paul Sabatier. Lahdili was supported by a postdoctoral fellowship of the BICMR in Beijing. We are very grateful to the referee for their careful reading of the manuscript and many valuable suggestions, which substantially improved our work. We thank D Calderbank, T Darvas, R Dervan, E Inoue, E Legendre, C Li and G Tian for their interest and comments on the manuscript. Apostolov thanks P Gauduchon for sharing with him his unpublished notes, and C Tønnesen-Friedman for bringing [42] to our attention.

1 Preliminaries on the weighted cscK problem

We recall the setup from [55]. Let X be a smooth compact, connected Kähler manifold of (real) dimension $2m$, and let

$$\mathcal{K}(X,\omega_0) = \{\varphi \in C^\infty(X) \mid \omega_\varphi := \omega_0 + dd^c\varphi > 0\}$$

be the space of ω_0–relative smooth Kähler potentials on X. We let $\mathbb{T} \subset \mathrm{Aut}_r(X)$ be a fixed compact torus in the *reduced* group of automorphisms of X, ie the connected closed subgroup $\mathrm{Aut}_r(X)$ of the group of complex automorphisms $\mathrm{Aut}(X)$, whose Lie algebra is the space of holomorphic vector fields of X with zeros; see eg [41]. Equivalently, $\mathrm{Aut}_r(X)$ is the connected component of the identity of the kernel of the natural group homomorphism from $\mathrm{Aut}(X)$ to the Albanese torus, and is known to be isomorphic to the linear algebraic group in the Chevalley-type decomposition of $\mathrm{Aut}(X)$; see [38]. We denote by $C^\infty_{\mathbb{T}}(X)$ the space of $\mathbb{T}$–invariant smooth functions on X and introduce the space

$$\mathcal{K}_{\mathbb{T}}(X,\omega_0) := \mathcal{K}(X,\omega_0) \cap C^\infty_{\mathbb{T}}(X),$$

of $\mathbb{T}$–invariant relative Kähler potentials, assuming also that ω_0 is $\mathbb{T}$–invariant.

It is well known that the action of $\mathbb{T}$ on (X,ω_0) is hamiltonian, and we let $\mu_0\colon X \to \mathfrak{t}^*$ be a momentum map, where $\mathfrak{t}$ is the Lie algebra of $\mathbb{T}$ and $\mathfrak{t}^*$ the dual vector space. By the convexity theorem [9; 44], the image $\Delta := \mu_0(X) \subset \mathfrak{t}^*$ is a compact convex polytope. For any $\varphi \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$, the smooth $\mathfrak{t}^*$–valued function

$$\mu_\varphi = \mu_0 + d^c\varphi \tag{5}$$

is the $\mathbb{T}$–momentum map of (X,ω_φ), normalized by the condition $\mu_\varphi(X) = \Delta$. In the above formula, $d^c\varphi$ is viewed as a smooth $\mathfrak{t}^*$–valued function via the identity $\langle d^c\varphi, \xi\rangle := d^c\varphi(\xi)$ for any $\xi \in \mathfrak{t} \subset C^\infty(X, TX)$.

1.1 The (v, w)–constant scalar curvature Kähler metrics

Following [55], let $v(\mu) > 0$ and $w(\mu)$ be smooth functions on Δ. One can then consider the condition (1) for a $\mathbb{T}$–invariant Kähler metric ω_φ in α (and the fixed polytope Δ), called a (v,w)*–cscK metric*. We thus want to solve the PDE for $\varphi \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$

$$\mathrm{Scal}_v(\omega_\varphi) = w(\mu_\varphi), \tag{6}$$

where

$$\mathrm{Scal}_v(\omega_\varphi) := v(\mu_\varphi)\,\mathrm{Scal}(\omega_\varphi) + 2\Delta_{\omega_\varphi} v(\mu_\varphi) + \langle g_\varphi, \mu_\varphi^*(\mathrm{Hess}(v))\rangle.$$

As we explained in the introduction, the problem of finding $\omega_\varphi \in \alpha$ solving (6) is obstructed by the condition (3), and in the case when v and w_0 are positive weights, this can be resolved (similarly to the approach in [21]) by finding a unique affine-linear function $\ell^{\mathrm{ext}}_{v,w_0}(\mu)$ on $\mathfrak{t}^*$, called *the extremal function*, such that for any ω_φ,

$$\int_X (\mathrm{Scal}_v(\omega_\varphi) - \ell^{\mathrm{ext}}_{v,w_0}(\mu_\varphi) w_0(\mu_\varphi))\ell(\mu_\varphi)\omega_\varphi^{[m]} = 0 \quad \text{for all } \ell \in \mathrm{Aff}(\mathfrak{t}^*).$$

Geometrically, the above condition means that the weighted cscK problem with weights $(v, w) = (v, \ell^{\mathrm{ext}}_{v,w_0} w_0)$ is unobstructed in terms of (3), and a solution ω_φ of the $(v, \ell^{\mathrm{ext}}_{v,w_0} w_0)$–cscK problem is referred to as a (v, w_0)*–extremal metric*.

1.2 The weighted Mabuchi energy

Definition 1.1 [55] Let v and w be weight functions on Δ with $v(\mu) > 0$. The weighted Mabuchi energy $\boldsymbol{M}_{v,w}$ on $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$ is defined by

$$(d_\varphi \boldsymbol{M}_{v,w})(\dot\varphi) = -\int_X (\mathrm{Scal}_v(\omega_\varphi) - w(\mu_\varphi))\dot\varphi\,\omega_\varphi^{[m]}, \quad \boldsymbol{M}_{v,w}(0) = 0.$$

Remark 1.2 It follows from the above definition and the results in [55] that for a constant c, $\boldsymbol{M}_{v,w}(\varphi + c) = \boldsymbol{M}_{v,w}(\varphi)$ if and only if v and w satisfy the integral relation

$$\int_X \mathrm{Scal}_v(\omega_0)\omega_0^{[m]} = \int_X w(\mu_0)\omega_0^{[m]}. \tag{7}$$

Furthermore, by the results in [55], (7) is a necessary condition for the existence of a solution of (6) and it is incorporated in the definition of $\boldsymbol{M}_{v,w}$ given in [55] via the constant $c_{v,w}(\alpha)$ in front of w, but we do not assume a priori this condition in the current article. It is however automatically satisfied if α admits a $\mathbb{T}$–invariant (v, w)–cscK metric, or if we consider the weights $(v, w) = (v, \ell^{\mathrm{ext}}_{v,w_0} w_0)$ corresponding to (v, w_0)–extremal Kähler metrics. In these cases, we shall write $\boldsymbol{M}_{v,w}(\omega_\varphi)$ to emphasize that the weighted Mabuchi functional acts on the space of $\mathbb{T}$–invariant Kähler metrics in $\alpha = [\omega_0]$.

The following result is established in [56], generalizing [11] to arbitrary weights $v > 0$ and w:

Theorem 1.3 *If ω is a $\mathbb{T}$–invariant (v, w)–cscK metric on $(X, \alpha, \mathbb{T}, \Delta)$, then for any $\varphi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ we have $\boldsymbol{M}_{v,w}(\omega_\varphi) \geq \boldsymbol{M}_{v,w}(\omega)$.*

1.3 The automorphism group of a (v, w_0)–extremal Kähler manifold

In what follows we will consider *connected* Lie groups. We recall that we have set $\mathrm{Aut}_r(X)$ to be the connected component of the identity of the kernel of the Albanese homomorphism and, similarly, we denote by $\mathrm{Aut}_r^{\mathbb{T}}(X)$ the connected component of the identity of the centralizer of the torus $\mathbb{T}$ in $\mathrm{Aut}_r(X)$. We shall use the following result, established in [55, Theorem B.1] (see also [39]) and [56, Remark 2]:

Proposition 1.4 *If $(X, \alpha, \mathbb{T})$ admits a (v, w_0)–extremal Kähler metric ω, then the connected component of the identity $\mathrm{Aut}_r^{\mathbb{T}}(X)$ of the subgroup of $\mathbb{T}$–commuting automorphisms in $\mathrm{Aut}_r(X)$ is reductive, and ω is invariant under the action of a maximal compact connected subgroup of $\mathrm{Aut}_r^{\mathbb{T}}(X)$. In particular, the isometry group of (X, ω) contains a maximal torus $\mathbb{T}_{\max} \subset \mathrm{Aut}_r(X)$ with $\mathbb{T} \subset \mathbb{T}_{\max}$. If, furthermore, $\mathbb{T} = \mathbb{T}_{\max}$, then $\mathrm{Aut}_r^{\mathbb{T}}(X) = \mathbb{T}^{\mathbb{C}}$.*

Because of this result, we shall often assume (without loss of generality for solving (6)) that $\mathbb{T} = \mathbb{T}_{\max} \subset \mathrm{Aut}_r(X)$ and thus $\mathrm{Aut}_r^{\mathbb{T}}(X) = \mathbb{T}^{\mathbb{C}}$.

1.4 Uniqueness of the (v, w_0)–extremal Kähler metrics

Another key result in the theory is the extension in [56] of the uniqueness results [11; 24] to the weighted setting.

Theorem 1.5 *Suppose ω and ω' are $\mathbb{T}$–invariant (v, w_0)–extremal Kähler metrics. Then there exists $\sigma \in \mathrm{Aut}_r^{\mathbb{T}}(X)$ such that $\sigma^*(\omega') = \omega$. In particular, if $\mathbb{T} \subset \mathrm{Aut}_r(X)$ is maximal, then the uniqueness holds modulo $\mathbb{T}^{\mathbb{C}}$.*

2 v–solitons as weighted cscK metrics

We review here the definition of v–solitons on a Fano manifold, following [15; 45], and discuss their link with (v, w)–cscK metrics.

We thus suppose throughout this section that X is a smooth Fano manifold $\alpha := 2\pi c_1(X)$ and $\mathbb{T} \subset \mathrm{Aut}(X)$ a fixed compact torus. (We recall here that on a Fano manifold $\mathrm{Aut}_r(X)$ coincides with the connected component of the identity of the full automorphism group.) We further consider the *natural* action of $\mathbb{T}$ on the anticanonical bundle K_X^{-1} of X, which normalizes the momentum map μ_ω of each $\mathbb{T}$–invariant Kähler metric $\omega \in \alpha$, and fixes the momentum image Δ. We shall sometimes refer to this normalization as the *canonical normalization* of Δ. In this setup, we recall:

Definition 2.1 Let $v > 0$ be a positive smooth weight function on Δ. A v–soliton on X is a $\mathbb{T}$–invariant Kähler metric $\omega \in 2\pi c_1(X)$ which satisfies the relation (4).

In the special case $v = e^{\langle \xi,\mu\rangle}$ we obtain a Kähler–Ricci soliton in the sense of [71].

Lemma 2.2 *A $\mathbb{T}$–invariant Kähler metric $\omega \in 2\pi c_1(X)$ is a v–soliton if and only if ω is a (v,w)–cscK metric with weight $w(\mu) = 2v(\mu)[m + \langle d\log v(\mu), \mu\rangle]$.*

Proof We start by showing that (4) implies that ω is (v,w)–cscK with the weight w specified in the lemma. Taking the trace in (4) with respect to ω gives

$$\begin{aligned}
\text{(8)}\quad \mathrm{Scal}(\omega) - 2m &= -\Delta_\omega(\log v(\mu_\omega)) \\
&= -\frac{1}{v(\mu_\omega)}\Delta_\omega(v(\mu_\omega)) - \frac{1}{v(\mu_\omega)^2} g_\omega(dv(\mu_\omega), dv(\mu_\omega)) \\
&= -\sum_{i=1}^m \frac{v_{,i}(\mu_\omega)}{v(\mu_\omega)}(\Delta_\omega \mu_\omega^{\xi_i}) + \sum_{i,j=1}^m \frac{v_{,ij}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j) \\
&\qquad - \sum_{i,j=1}^m \frac{v_{,i}(\mu_\omega)v_{,j}(\mu_\omega)}{v(\mu_\omega)^2} g_\omega(\xi_i,\xi_j),
\end{aligned}$$

where $(\xi_i)_{i=1,\dots,r}$ is a basis of $\mathfrak{t}$ and $v_{,i}$ denotes the partial derivative in direction of ξ_i. On the other hand, by taking the interior product of (4) with ξ_i and using that ξ_i is Killing with respect to ω, we get

$$-d\Delta_\omega \mu_\omega^{\xi_i} + 2d\mu_\omega^{\xi_i} = d\big(d^c(\log v(\mu_\omega))(\xi_i)\big) = d\bigg(\sum_{j=1}^m \frac{v_{,j}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j)_\omega\bigg),$$

where $\mu_\omega^\xi := \langle \mu_\omega, \xi\rangle$ is the momentum of ξ. It follows that

$$-\Delta_\omega \mu_\omega^{\xi_i} + 2\mu_\omega^{\xi_i} = \sum_{j=1}^m \frac{v_{,j}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j) + c \tag{9}$$

for some constant c. As we consider the canonical normalization of μ_ω (corresponding to the natural lifted $\mathbb{T}$–action on K_X^{-1}), one can see that $c = 0$. Indeed, the infinitesimal actions A_i of the elements of the basis $(\xi_i)_i$ on smooth sections of K_X^{-1} are given by $A_i(s) := \mathcal{L}_{\xi_i} s$. We denote by H_g the induced hermitian metric on K_X^{-1} through the riemannian metric g_ω of ω (so that H_g has curvature ρ_ω) and by $H = v(\mu_\omega)H_g$ the induced hermitian metric with curvature ω (by using (4)); comparing the actions of the corresponding Chern connections, $\nabla^g_{\xi_i}$ and $\nabla^H_{\xi_i} = \nabla^g_{\xi_i} - \frac{1}{2}\sqrt{-1}d^c \log v(\mu_\omega)(\xi_i)\,\mathrm{id}$ on smooth sections of K_X^{-1} with the infinitesimal actions A_i gives (see eg [41, Propositions 8.8.2 and 8.8.3])

$$A_i(s) = \nabla^g_{\xi_i} s + \tfrac{1}{2}\sqrt{-1}(\Delta_\omega \mu_\omega^{\xi_i})s \quad \text{and} \quad A_i(s) = \nabla^H_{\xi_i} s + \sqrt{-1}\mu_\omega^{\xi_i} s. \tag{10}$$

We thus deduce $\frac{1}{2}\Delta_\omega \mu_\omega^{\xi_i} = \mu_\omega^{\xi_i} - \frac{1}{2}d^c(\log v(\mu_\omega))(\xi_i)$, ie $c = 0$ in (9).

Now, letting $c = 0$ in (9), multiplying it by $v_{,i}(\mu_\omega)/v(\mu_\omega)$, and taking the sum over i gives

$$\sum_{i,j=1}^m \frac{v_{,i}(\mu_\omega)v_{,j}(\mu_\omega)}{v(\mu_\omega)^2} g_\omega(\xi_i,\xi_j) = \sum_{i=1}^r \frac{v_{,i}(\mu_\omega)}{v(\mu_\omega)}(\Delta_\omega \mu_\omega^{\xi_i}) - 2\sum_{i=1}^r \frac{v_{,i}(\mu_\omega)}{v(\mu_\omega)}\mu_\omega^{\xi_i},$$

which, substituted back into (8), yields

$$\begin{aligned}
\mathrm{Scal}(\omega) - 2m &= -2\sum_{i=1}^r \frac{v_{,i}(\mu_\omega)}{v(\mu_\omega)}\Delta_\omega \mu_\omega^{\xi_i} + \sum_{i,j=1}^m \frac{v_{,ij}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j) + 2\sum_{i=1}^r \frac{v_{,i}(\mu_\omega)}{v(\mu_\omega)}\mu_\omega^{\xi_i} \\
&= -2\sum_{i=1}^r \frac{v_{,i}(\mu_\omega)}{v(\mu_\omega)}\Delta_\omega \mu_\omega^{\xi_i} + 2\sum_{i,j=1}^m \frac{v_{,ij}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j) \\
&\qquad\qquad - \sum_{i,j=1}^m \frac{v_{,ij}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j) + 2\langle d\log v, \mu_\omega\rangle \\
&= -2\Delta_\omega(v(\mu_\omega)) - \sum_{i,j=1}^m \frac{v_{,ij}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j) + 2\langle d\log v, \mu_\omega\rangle.
\end{aligned}$$

Thus, $\mathrm{Scal}_v(\omega) = w(\mu_\omega)$.

Now we show the converse. To this end, let $\omega \in 2\pi c_1(X)$ be a $\mathbb{T}$–invariant Kähler metric, $v > 0$ a positive smooth function on the canonically normalized polytope Δ and $w = 2(m + \langle d\log, \mu\rangle)v$ the weight defined in Lemma 2.2. Let $h \in C^\infty_{\mathbb{T}}(X)$ be an ω–relative Ricci potential, ie

$$\rho_\omega - \omega = \tfrac{1}{2}dd^c h.$$

Taking the trace with respect to ω and the interior product with $\xi \in \mathfrak{t}$ in the above identity, we get

$$\mathrm{Scal}(\omega) = 2m - \Delta_\omega h \quad \text{and} \quad \Delta_\omega \mu_\omega^\xi + \mathcal{L}_{J\xi} h = 2\mu_\omega^\xi, \tag{11}$$

where we have used the canonical normalization of μ_ω to determine the additive constant in the second inequality (as we did for (9)). Similar computations as in the first part of the proof (using (11)) give

$$\begin{aligned}
(12)\quad \operatorname{Scal}_v(\omega) - w(\mu_\omega) &= -v(\mu_\omega)(\Delta_\omega h) + 2\sum_{i=1}^r v_{,i}(\mu_\omega)(\Delta_\omega \mu_\omega^{\xi_i}) - \sum_{i,j=1}^r v_{,ij}(\mu_\omega)(\xi_i,\xi_j)_\omega \\
&\qquad + 2\sum_{i=1}^r v_{,i}(\mu_\omega)\mu_\omega^{\xi_i} \\
&= -v(\mu_\omega)(\Delta_\omega h) + \sum_{i=1}^r v_{,i}(\mu_\omega) g_\omega(dh, d\mu_\omega^{\xi_i}) + \sum_{i=1}^r v_{,i}(\mu_\omega)(\Delta_\omega \mu_\omega^{\xi_i}) \\
&\qquad - \sum_{i,j=1}^r v_{,ij}(\mu_\omega) g_\omega(\xi_i,\xi_j) \\
&= -v(\mu_\omega)(\Delta_{\omega,v} h) + \sum_{i=1}^r v_{,i}(\mu_\omega)(\Delta_\omega \mu_\omega^{\xi_i}) - \sum_{i,j=1}^r v_{,ij}(\mu_\omega) g_\omega(\xi_i,\xi_j),
\end{aligned}$$

where $\Delta_{\omega,v} := (1/v(\mu_\omega))\delta_\omega v(\mu_\omega) d$ is the weighted Laplacian; see Appendix A. Using the second equality in (8), we compute

$$\begin{aligned}
v(\mu_\omega)\Delta_{\omega,v}(\log v(\mu_\omega)) &= v(\mu_\omega)(\Delta_\omega \log v(\mu_\omega)) - \sum_{i=1}^m v_i(\mu_\omega) g_\omega(d(\log v(\mu_\omega)), d\mu_\omega^{\xi_i}) \\
&= v(\mu_\omega)(\Delta_\omega \log v(\mu_\omega)) - \sum_{i,j=1}^m \frac{v_{,i}(\mu_\omega) v_{,j}(\mu_\omega)}{v(\mu_\omega)} g_\omega(\xi_i,\xi_j) \\
&= \sum_{i=1}^r v_{,i}(\mu_\omega)(\Delta_\omega \mu_\omega^{\xi_i}) - \sum_{i,j=1}^r v_{,ij}(\mu_\omega) g_\omega(\xi_i,\xi_j).
\end{aligned}$$

Substituting back in (12),

$$\operatorname{Scal}_v(\omega) - w(\mu_\omega) = v(\mu_\omega)\Delta_{\omega,v}(\log v(\mu_\omega) - h). \tag{13}$$

It follows that if ω is (v,w)–cscK then $h = \log v(\mu_\omega) + c$ by the maximum principle, showing that ω satisfies (4). □

Remark 2.3 Using the second relation in (11) it follows that, under the canonical normalization of μ_ω,

$$\int_X \mu_\omega^\xi e^h \omega^{[m]} = 0 \quad \text{for } \xi \in \mathfrak{t}. \tag{14}$$

This is precisely the normalization of μ_ω used in [71, Section 2].

Lemma 2.4 *Define $v := \ell^{-(m+2)}$ for $\ell(\mu) = \langle \xi, \mu \rangle + a$ positive affine-linear on Δ. Then $\omega \in 2\pi c_1(X)$ is a v–soliton if and only if ω is an $(\ell^{-(m+1)}, 2ma\ell^{-(m+2)})$–cscK metric.*

Proof The proof is similar to that of Lemma 2.2.

If ω is a v–soliton with $v := \ell^{-(m+2)}$, specializing (8) and (9) to the specific choice of v and letting $f := \ell(\mu_\omega) = \mu_\omega^\xi + a$, we get the identities

$$\mathrm{Scal}(\omega) = 2m + (m+2)\Delta_\omega \log f \quad \text{and} \quad -\Delta_\omega f + 2f = \frac{m+2}{f} g_\omega(df, df) + 2a.$$

Multiplying the first equality by f^2 and taking the sum with the second equality multiplied by mf gives

$$f^2\,\mathrm{Scal}(\omega) - 2(m+1) f \Delta_\omega f - (m+1)(m+2) g_\omega(df, df) = 2maf. \tag{15}$$

The right side is the $(m+2, f)$–scalar curvature (see [2]) and it is straightforward to check that (15) is equivalent to the condition that ω is an $(\ell^{-(m+1)}, 2ma\ell^{-(m+2]})$–cscK metric.

In the other direction, for any $\mathbb{T}$–invariant Kähler metric $\omega \in 2\pi c_1(X)$, we let

$$f := \ell(\mu_\omega) = \mu_\omega^\xi + a > 0$$

be the corresponding Killing potential and let $h \in C^\infty_{\mathbb{T}}(X)$ be such that $\rho_\omega - \omega = \frac{1}{2} dd^c h$. From (11) we have

$$\mathrm{Scal}(\omega) = 2m - \Delta_\omega h \quad \text{and} \quad -\Delta_\omega f + 2f = -g_\omega(df, dh) + 2a.$$

Multiplying the first identity by f^2 and summing with the second identity multiplied by mf gives

$$\begin{aligned} f^2\,\mathrm{Scal}(\omega) - 2(m+1) f \Delta_\omega f - (m+1)(m+2) g_\omega(df, df) - 2maf \\ = -f^2\big(\Delta_\omega(h + (m+2)\log f) + m g_\omega(d\log f, dh + (m+2)\, d\log f)\big). \end{aligned} \tag{16}$$

If we suppose that (15) holds, we conclude, again by the maximum principle, that $(m+2)\log f + h$ must be constant. □

Remark 2.5 Lemmas 2.2 and 2.4 give two different realizations of the same $\ell^{-(m+2)}$–soliton as a weighted cscK metric, with weights $\big(\ell^{-(m+2)}, 2(-2\ell + (m+2)a)\ell^{-(m+3)}\big)$ and $(\ell^{-(m+1)}, 2am\ell^{-(m+2)})$, respectively.

We derive from Lemma 2.4 and the correspondence in [2] the following fact, which does not seem to have been noticed before:

Lemma 2.6 *On a Fano manifold $(X,\mathbb{T})$, a $\mathbb{T}$–invariant Kähler metric $\omega \in 2\pi c_1(X)$ is an $\ell^{-(m+2)}$–soliton with respect to a positive affine-linear function $\ell = \langle \xi, \mu\rangle + a$ if and only if the lift $\hat{\xi}$ of the vector field ξ to K_X, via the hermitian connection ∇^h with curvature $-\omega$ and the ω–momentum $\ell(\mu_\omega)$ of ξ, is a Reeb vector of a Sasaki–Einstein (transversal) structure of transversal scalar curvature $2am$, defined on the unit circle bundle N of (K_X, h).*

Proof By Lemma 2.4, we need to show that an $(\ell^{-(m+1)}, 2am\ell^{-(m+2)})$–cscK metric in $2\pi c_1(X)$ corresponds to a Sasaki–Einstein structure as defined in the statement. By [2, Theorem 1], the condition that ω is $(\ell^{-(m+1)}, 2am\ell^{-(m+2)})$–cscK is equivalent to the condition that the corresponding Sasaki structure has transversal scalar curvature equal to $2ma$ (notice that $a > 0$ by the positivity of ℓ over the canonical polytope Δ). Any Sasaki structure of constant transversal scalar curvature on $N \subset K_X$ is transversally Kähler–Einstein as $c_1(K_X^\times) = 0$, and therefore the first Chern class of the CR distribution of N vanishes; see eg [19, Corollary 5.3; 40, Proposition 4.3]. □

Remark 2.7 The correspondence in Lemma 2.6 is, in fact, local and can be deduced directly from the relation between the transversal Ricci tensors of the two Sasaki structures on the CR manifold $N \subset K_X$, defined by $\hat{\xi}$ and the regular Reeb vector field $\hat{\chi}$, respectively, according to [51] and Gauduchon (personal communication).

Proof of Propositions 1 and 2 Propositions 1 and 2 from the introduction follow directly from Lemmas 2.2, 2.4 and 2.6. □

3 The coercivity principle: the plan of the proof of Theorem 1

We consider the following general setup, based on the results of [27; 29; 70]. As before, we let $\mathbb{T} \subset \mathrm{Aut}_r(X)$ be a fixed connected compact torus in the reduced group of automorphisms of X, and denote by $\mathbb{G} = \mathbb{T}^{\mathbb{C}} \subset \mathrm{Aut}_r(X)$ the corresponding complex torus.

Following [27], we consider the L_1–length function on $\mathcal{K}(X,\omega_0)$, introduced on a smooth curve ψ_t for $t \in [0,1]$ by

$$L_1(\psi_t) := \int_0^1 \Big(\int_X |\dot{\psi}_t|\omega_{\psi_t}^{[m]}\Big)\, ds,$$

and, for $\varphi_0, \varphi_1 \in \mathcal{K}(X,\omega_0)$, we let

$$d_1(\varphi_0,\varphi_1) := \inf\{L_1(\psi_t) \mid \psi_t \in \mathcal{K}(X,\omega_0) \text{ for } t \in [0,1], \psi_0 = \varphi_0 \text{ and } \psi_1 = \varphi_1\}.$$

Similarly, we define d_1 on $\mathcal{K}_{\mathbb{T}}(M,\omega_0)$ by considering the infimum over smooth curves in $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$. It is proved in [27] that $(\mathcal{K}(X,\omega_0),d_1)$ is a metric space, and it is observed in [29] that $(\mathcal{K}_{\mathbb{T}}(X,\omega_0),d_1)$ is a metric subspace of $(\mathcal{K}(X,\omega_0),d_1)$.

Recall the following well-known functionals on $\mathcal{K}(X,\omega_0)$:

Definition 3.1 Let $\boldsymbol{I}$ denote the functional on $\mathcal{K}(X,\omega_0)$ defined by

$$(d_\varphi \boldsymbol{I})(\dot\varphi)=\int_X \dot\varphi\,\omega_\varphi^{[m]},\quad \boldsymbol{I}(0)=0,$$

and let $\boldsymbol{J}(\varphi):=\int_X \varphi\omega_0^{[m]}-\boldsymbol{I}(\varphi)$.

Remark 3.2 For any constant c, we have that $\boldsymbol{I}(\varphi+c)=\boldsymbol{I}(\varphi)+c\,\mathrm{Vol}(X,\omega_0)$ (where $\mathrm{Vol}(X,\omega_0)=\int_X {\omega_0}^{[m]}$ is the total volume of (X,ω_0)), whereas $\boldsymbol{J}(\varphi+c)=\boldsymbol{J}(\varphi)$, ie we can see $\boldsymbol{J}$ as a functional on the space of Kähler metrics in the Kähler class $\alpha=[\omega_0]$, which motivates the notation $\boldsymbol{J}(\omega_\varphi)$. One can further show that $\boldsymbol{J}(\omega_\varphi)\geq 0$ with equality if and only if $\omega_\varphi=\omega_0$.

By the above remark, for any Kähler metric ω_φ in the Kähler class $[\omega_0]$, there exists a uniquely determined ω_0–relative potential $\varphi\in\mathcal{K}(X,\omega_0)$ satisfying

$$\boldsymbol{I}(\varphi)=0.$$

We shall denote by $\mathring{\mathcal{K}}(X,\omega_0)$ (and $\mathring{\mathcal{K}}_{\mathbb{T}}(X,\omega_0)$) the subspaces of *normalized* ω_0–relative Kähler potentials satisfying the above equality. We notice that the group $\mathbb{G}=\mathbb{T}^{\mathbb{C}}$ naturally acts on the space of Kähler metrics in $[\omega_0]$, preserving the subspace of $\mathbb{T}$–invariant Kähler metrics. This induces an action $[\mathbb{G}]$ on the spaces $\mathring{\mathcal{K}}(X,\omega_0)$ and $\mathring{\mathcal{K}}_{\mathbb{T}}(X,\omega_0)$ such that

$$\omega_{\sigma[\varphi]}=\sigma^*(\omega_\varphi)\quad\text{for all } \sigma\in\mathbb{G} \text{ and } \varphi\in\mathring{\mathcal{K}}(X,\omega_0).$$

We introduce the $\mathbb{G}$–relative distance on $\mathring{\mathcal{K}}(X,\omega_0)$ and $\mathring{\mathcal{K}}_{\mathbb{T}}(X,\omega_0)$ by

$$d_1^{[\mathbb{G}]}(\varphi_0,\varphi_1)=\inf_{\sigma_0,\sigma_1\in\mathbb{G}} d_1(\sigma_0[\varphi_0],\sigma_1[\varphi_1]).$$

It is proved in [29] that $d_1^{[\mathbb{G}]}$ is $\mathbb{G}$–invariant, ie $d_1^{[\mathbb{G}]}(\sigma[\varphi_0],\sigma[\varphi_1])=d_1^{[\mathbb{G}]}(\varphi_0,\varphi_1)$, and thus

$$d_1^{[\mathbb{G}]}(\varphi_0,\varphi_1)=\inf_{\sigma\in\mathbb{G}} d_1(\varphi_0,\sigma[\varphi_1]).$$

Definition 3.3 Let $\boldsymbol{F}$ be a functional on $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$. We say that $\boldsymbol{F}$ is $\mathbb{G}$–*coercive* if there exist uniform positive constants (λ,δ) such that

$$\boldsymbol{F}(\varphi)\geq\lambda d_1^{[\mathbb{G}]}(0,\varphi)-\delta\quad\text{for all } \varphi\in\mathring{\mathcal{K}}_{\mathbb{T}}(X,\omega_0). \tag{17}$$

It is sometimes more natural to introduce $\mathbb{G}$–coercivity in terms of the functional $\boldsymbol{J}$ via:

Proposition 3.4 [29] *$\boldsymbol{F}$ is $\mathbb{G}$–coercive if and only if there exist uniform positive constants (λ', δ') such that*

$$\boldsymbol{F}(\varphi) \geq \lambda' \inf_{\sigma \in \mathbb{G}} \boldsymbol{J}(\sigma^* \omega_\varphi) - \delta' \quad \text{for all } \varphi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0). \tag{18}$$

Remark 3.5 If $\boldsymbol{F}$ is $\mathbb{G}$–coercive, then it is bounded below by (17).

As in [27], one can consider the metric completion $(\mathcal{E}^1(X, \omega_0), d_1)$ of $(\mathcal{K}(X, \omega_0), d_1)$, which can be characterized by a suitable continuously embedded subspace in $L^1(X, \omega_0)$; similarly we let $(\mathcal{E}^1_{\mathbb{T}}(X, \omega_0), d_1)$ be the metric completion of $(\mathcal{K}_{\mathbb{T}}(X, \omega_0), d_1)$, which, again by the results in [29], can be viewed as the closed subspace of $\mathbb{T}$–invariant elements of $\mathcal{E}^1(X, \omega_0)$. It will be important for us that $(\mathcal{E}^1_{\mathbb{T}}(X, \omega_0), d_1)$ is a *geodesic space*, ie each two elements $\psi_0, \psi_1 \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ can be connected with a curve ψ_t for $t \in [0, 1]$ in $(\mathcal{E}^1_{\mathbb{T}}(X, \omega_0), d_1)$, called a *weak geodesic*, obtained as the limit of $C^{1,\bar{1}}$–geodesics between elements of $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$; see [22; 27]. This object is a curve $\varphi_t \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$, of regularity $C^{1,1}([0, 1] \times X)$, which is uniquely associated to each $\varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$; see [16; 22; 25] and the proof of Proposition 5.8 for more details about the weak $C^{1,\bar{1}}$–geodesics.

In [29, Theorem 3.4], the following general principle is established:

Theorem 3.6 (coercivity principle) *Let $\boldsymbol{F} : \mathcal{K}_{\mathbb{T}}(X, \omega_0) \to \mathbb{R}$ be a lower semicontinuous (lsc) functional with respect to d_1, and $\boldsymbol{F} : \mathcal{E}^1_{\mathbb{T}}(X, \omega_0) \to \mathbb{R} \cup \{+\infty\}$ be its largest lsc extension. Suppose, furthermore, that $\boldsymbol{F}(\varphi + c) = \boldsymbol{F}(\varphi) =: \boldsymbol{F}(\omega_\varphi)$ and $\boldsymbol{F}(\sigma^* \omega_\varphi) = \boldsymbol{F}(\omega_\varphi)$ for any $\varphi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ and $\sigma \in \mathbb{G}$, and that $\boldsymbol{F}$ satisfies:*

- (i) ***Convexity*** *For each $\varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ and the $C^{1,\bar{1}}$–geodesic φ_t joining φ_0 and φ_1, $t \to \boldsymbol{F}(\varphi_t)$ is continuous and convex.*
- (ii) ***Regularity*** *If $\psi \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ is a minimizer of $\boldsymbol{F}$, then $\psi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$.*
- (iii) ***Uniqueness*** *$\mathbb{G}$ acts transitively on the set of minimizers of $\boldsymbol{F}$.*
- (iv) ***Compactness*** *If $\{\psi_j\}_j \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ satisfies $\lim_{j \to \infty} \boldsymbol{F}(\psi_j) = \inf_{\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)} \boldsymbol{F}$ and, for some $C > 0$, $d_1(0, \psi_j) \leq C$, then there exists a $\psi \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ and a subsequence $\{\psi_{j_k}\}_k$ with $\psi_{j_k} \to \psi$ in $(\mathcal{E}^1_{\mathbb{T}}(X, \omega_0), d_1)$.*

Then, the following two conditions are equivalent:

- *$\boldsymbol{F}$ has minimizer in $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$.*
- *$\boldsymbol{F}$ is $\mathbb{G}$**–coercive**.*

The above result provides a clear framework for achieving the proof of Theorem 1: we need to find a suitable largest lsc extension of the weighted Mabuchi functional $\boldsymbol{M}_{v,w}$ to the space $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ and show it satisfies the properties (i)–(iv). Notice that the invariance of $\boldsymbol{M}_{v,w}$ under the action of $\mathbb{G}=\mathbb{T}^{\mathbb{C}}$ is equivalent to the necessary condition (3) for the existence of a (v,w)–cscK metric, whereas (iii) will follow from Theorem 1.5 once the regularity condition (ii) is established. Furthermore, the property (i) is proved in [56, Theorem 1], so the core of our argument is to define the extension of $\boldsymbol{M}_{v,w}$ to $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ and establish the properties (ii) and (iv). These steps will be detailed in Theorems 6.1, 7.1 and 6.17, respectively.

4 K–stability via coercivity: deriving Corollary 1 from Theorem 1

We use the following general setup, based on the results of [10; 14; 18; 49; 57; 66; 70] which deal with the K–polystability and uniform K–stability in the unweighted cscK case. Let $\mathbb{T}\subset \mathrm{Aut}_r(X)$ be a connected compact torus in the reduced group of automorphisms of X.

Definition 4.1 A $\mathbb{T}$–equivariant Kähler test configuration $(\mathscr{X},\mathscr{A})$ associated to $(X,\alpha,\mathbb{T})$ is a *normal* compact Kähler space $\mathscr{X}$ endowed with

- a flat morphism $\pi\colon \mathscr{X}\to\mathbb{P}^1$;
- a $\mathbb{C}^*$–action ρ covering the standard $\mathbb{C}^*$–action on $\mathbb{P}^1$, and a $\mathbb{T}$–action commuting with ρ and preserving π;
- a $\mathbb{T}\times\mathbb{C}^*$–equivariant biholomorphism $\Pi_0\colon(\mathscr{X},\setminus\pi^{-1}(0))\cong X\times(\mathbb{P}^1\setminus\{0\})$;
- a Kähler class $\mathscr{A}\in H^{1,1}(\mathscr{X},\mathbb{R})$ such that $(\Pi_0^{-1})^*(\mathscr{A})|_{X\times\{\tau\}}=\alpha$.

We say that $(\mathscr{X},\mathscr{A})$ is *smooth* if $\mathscr{X}$ is smooth and *dominating* if Π_0 extends to a $\mathbb{T}\times\mathbb{C}^*$–equivariant morphism

$$\Pi\colon\mathscr{X}\to X\times\mathbb{P}^1. \tag{19}$$

$(\mathscr{X},\mathscr{A})$ is called *trivial* if it is dominating and Π is an isomorphism; $(\mathscr{X},\mathscr{A})$ is called *product* if $\pi^{-1}(0)\cong X$. If (X,L) is a smooth polarized variety and $\alpha=2\pi c_1(L)$, a *polarized* test configuration is a normal polarized variety $(\mathscr{X},\mathscr{L})$ such that, for some $r\in\mathbb{N}^*$, $(\mathscr{X},(1/r)2\pi c_1(\mathscr{L}))$ defines a Kähler test configuration of (X,α) and, under Π_0, $(X,\mathscr{L}|_{X\times\{\tau\}})\cong(X,L^r)$.

4.1 Non-Archimedean functionals

We recall that any $\mathbb{T}\times\mathbb{S}^1$–invariant Kähler metric $\Omega\in\mathscr{A}$ on $\mathscr{X}$ gives rise to a smooth ray of $\mathbb{T}$–invariant Kähler metrics $\omega_t\in\alpha$ on X defined by

$$\omega_t := \rho(e^{-t+is})^*(\Omega)|_{X\times\{1\}}.$$

Definition 4.2 Let $\boldsymbol{F}$ be a functional defined on the space of $\mathbb{T}$–invariant Kähler metrics on X in the class α. We say that $\boldsymbol{F}$ admits a *non-Archimedean version* $\boldsymbol{F}^{\mathrm{NA}}$, defined on a subclass C of $\mathbb{T}$–equivariant Kähler test configurations $(\mathscr{X},\mathscr{A})$ associated to $(X,\alpha,\mathbb{T})$ if, for any $(\mathscr{X},\mathscr{A})\in C$ and any induced smooth ray of $\mathbb{T}$–invariant Kähler metrics $\omega_t\in\alpha$ on X, the slope $\lim_{t\to\infty}\boldsymbol{F}(\omega_t)/t$ is well defined and given by a quantity $\boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A})$ which is independent of the choice of the $\mathbb{T}\times\mathbb{S}^1$–invariant Kähler form $\Omega\in\mathscr{A}$.

We give below two key examples of non-Archimedean versions of known functionals. The first one is established in the polarized case in [18] and in the generality we consider in [34; 65]:

Example 4.3 The functional $\boldsymbol{J}$ introduced in Definition 3.1 admits a non-Archimedean version defined, up to a positive-dimensional multiplicative constant, on the class of smooth $\mathbb{T}$–equivariant dominating Kähler test configurations $(\mathscr{X},\mathscr{L})$ by

$$\boldsymbol{J}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}) = \frac{((\Pi^*\alpha)^m\cdot\mathscr{A})_{\mathscr{X}}}{(\alpha^m)_X} - \frac{1}{m+1}\frac{(\mathscr{A}^{m+1})_{\mathscr{X}}}{(\alpha^m)_X},$$

where Π is the morphism (19) and α denotes both the Kähler class on X and its pullback to $X\times\mathbb{P}^1$.

The above expression generalizes to dominating smooth test configurations which are only *relatively nef* (in the terminology of [66]), thus also providing a non-Archimedean version of $\boldsymbol{J}$ for any Kähler test configuration. Indeed, by the equivariant Hironaka resolution, any $\mathbb{T}$–equivariant test configuration can be dominated by a smooth relatively nef Kähler dominating test configuration, and the computation of $\boldsymbol{J}^{\mathrm{NA}}$ on the latter does not depend on the choice made.

The non-Archimedean functional $\boldsymbol{J}^{\mathrm{NA}}$ defined above is always nonnegative and equals zero precisely when $(\mathscr{X},\mathscr{A})$ is the *trivial* test configuration. This statement is proved in [18, Theorem 7.9] in the polarized case, and follows from the results in [66] in the Kähler case; see in particular [66, Lemma 4.8] with G trivial and recall that the $\boldsymbol{J}$–norm is Lipschitz equivalent to the d_1–distance, so that the unique weak geodesic ray associated to a test configuration with vanishing $\boldsymbol{J}^{\mathrm{NA}}$–norm must be constant, and hence the test configuration must be trivial by [66, Corollary 3.12]. Thus, $\boldsymbol{J}^{\mathrm{NA}}$ can be thought of as a "norm" on the space of Kähler test configurations.

In order to obtain a norm which is zero for more general *product* test configurations, in [33; 49; 57] the authors consider smooth rays $\widetilde{\omega}_t \in \alpha$ of $\mathbb{T}$–invariant Kähler metrics on X which are obtained by composing an induced ray ω_t from a $\mathbb{T}\times\mathbb{S}^1$–invariant Kähler metric $\Omega \in \mathscr{A}$ on $\mathscr{X}$ with the flow of a vector field $J\xi$, where $\xi \in \mathfrak{t}$, ie $\widetilde{\omega}_t = \exp(tJ\xi)^*(\omega_t)$. They show that the slope

$$\lim_{t\to\infty} \frac{\boldsymbol{J}(\widetilde{\omega}_t)}{t} =: \boldsymbol{J}^{\mathrm{NA}}(\mathscr{X}_\xi, \mathscr{A}_\xi)$$

is well defined and independent of the choice of induced ray ω_t. We notice that when $\xi \in 2\pi\Lambda$ is a lattice element (or more generally is rational), ξ induces a $\mathbb{C}^*$–action ρ_ξ on $\mathscr{X}$ and $\widetilde{\omega}_t$ is an induced smooth ray from another Kähler test configuration $(\mathscr{X}_\xi, \mathscr{A}_\xi)$, called the ξ*–twist* of $(\mathscr{X}, \mathscr{A})$, obtained from $\mathscr{X}$ by composing the initial $\mathbb{C}^*$–action ρ with ρ_ξ and compactifying trivially at infinity. (For instance, the product test configurations are precisely the ξ–twists of the trivial test configuration.) In this case, $\boldsymbol{J}^{\mathrm{NA}}(\mathscr{X}_\xi, \mathscr{A}_\xi)$ is just the non-Archimedean $\boldsymbol{J}$–functional computed as in Example 4.3 on $(\mathscr{X}_\xi, \mathscr{A}_\xi)$. For a general ξ, the quantity $(\mathscr{X}_\xi, \mathscr{A}_\xi)$ in this notation is not a test configuration in the usual sense (it is sometimes refereed to as an $\mathbb{R}$*–test configuration*) but the value $\boldsymbol{J}^{\mathrm{NA}}(\mathscr{X}_\xi, \mathscr{A}_\xi)$ can be obtained as a continuous extension of the corresponding quantity for rational ξ's. Following [49; 57], we let

$$\boldsymbol{J}^{\mathrm{NA}}_{\mathbb{T}^{\mathbb{C}}}(\mathscr{X}, \mathscr{A}) := \inf_{\xi\in\mathfrak{t}} \boldsymbol{J}^{\mathrm{NA}}(\mathscr{X}_\xi, \mathscr{A}_\xi) \geq 0. \tag{20}$$

A key observation [18; 49; 57] in the polarized case is that the equality in (20) holds if and only if $(\mathscr{X}, \mathscr{L})$ is a *product* test configuration. Furthermore, according to [49, Theorem B; 57, Theorem 3.14]:

Example 4.3 (continued) In the polarized case, the quantity $\boldsymbol{J}^{\mathrm{NA}}_{\mathbb{T}^{\mathbb{C}}}(\mathscr{X}, \mathscr{A})$ introduced in (20) defines a non-Archimedean version of the functional

$$\boldsymbol{J}_{\mathbb{T}^{\mathbb{C}}}(\omega) := \inf_{\sigma\in\mathbb{T}^{\mathbb{C}}} \boldsymbol{J}(\sigma^*(\omega))$$

on the class of $\mathbb{T}$–equivariant polarized test configuration of $(X, L, \mathbb{T})$.

Our second example is established in [55, Theorem 7]:

Example 4.4 Consider the weighted Mabuchi functional $\boldsymbol{M}_{v,w}$ from Definition 1.1 and assume that the relation (7) holds; see Remark 1.2. Then $\boldsymbol{M}_{v,w}$ admits a non-Archimedean version defined on *smooth* $\mathbb{T}$–equivariant Kähler test configurations with *reduced central fibre*, given by

$$(21)\qquad \mathscr{F}_{v,w}(\mathscr{X},\mathscr{A}) := -\int_{\mathscr{X}} (\mathrm{Scal}_v(\Omega) - w(\mu_\Omega))\Omega^{[m+1]} + 8\pi \int_X v(\mu_\omega)\omega^{[m]},$$

where $\Omega \in \mathscr{A}$ is any $\mathbb{T}$–invariant Kähler metric on $\mathscr{X}$ with Δ–normalized $\mathbb{T}$–momentum map $\mu_\Omega\colon \mathscr{X} \to \Delta$ and v–scalar curvature $\mathrm{Scal}_v(\Omega)$, and $\omega \in \alpha$ is any $\mathbb{T}$–invariant Kähler metric on X with Δ–normalized $\mathbb{T}$–momentum map $\mu_\omega\colon X \to \Delta$.

Definition 4.5 The right side of (21) is independent of $\Omega \in \mathscr{A}$ and $\omega \in \alpha$ (see [55]) and is referred to as the (v,w)*–weighted Donaldson–Futaki invariant* of a smooth $\mathbb{T}$–equivariant Kähler test configuration $(\mathscr{X},\mathscr{A})$.

Remark 4.6 In the unweighted case (ie $v = 1$ and $w = 4m\pi c_1(X)\cdot\alpha^{m-1}/\alpha^m$), $\mathscr{F}_{v,w}(\mathscr{X},\mathscr{A})$ admits an equivalent expression in terms of intersection cohomology numbers on $\mathscr{X}$; see [62; 72]. This allows one to extend the definition of the (unweighted) Donaldson–Futaki invariant to *any* normal Kähler test configuration. For arbitrary weight functions $v > 0$ and w, we don't have as yet a general definition for $\mathscr{F}_{v,w}$, but (21) can be readily extended to orbifold test configurations. We also notice that the assumption on the central fibre in Example 4.4 is necessary in order to ensure the equality $\mathscr{F}_{v,w} = \boldsymbol{M}^{\mathrm{NA}}_{v,w}$; see [65] for a general formula of the non-Archimedean version of the unweighted Mabuchi energy. It will be interesting to obtain a non-Archimedean version of $\boldsymbol{M}_{v,w}$ for any orbifold $\mathbb{T}$–equivariant Kähler test configuration.

4.2 $\boldsymbol{F}^{\mathrm{NA}}$–stability

Definition 4.7 Let $\boldsymbol{F}$ be a functional defined on the space of $\mathbb{T}$–invariant Kähler metrics on X in the Kähler class α, and suppose $\boldsymbol{F}$ admits a non-Archimedean version $\boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A})$ (see Definition 4.2), defined on a class C of $\mathbb{T}$–equivariant Kähler test configurations $(\mathscr{X},\mathscr{A})$ associated to $(X,\alpha,\mathbb{T})$. We suppose that C contains the product test configurations. We say that:

(i) $(X,\alpha,\mathbb{T})$ is $\mathbb{T}$*–equivariant* $\boldsymbol{F}^{\mathrm{NA}}$*–semistable* (on test configurations of C) if for any $(\mathscr{X},\mathscr{A}) \in C$ we have $\boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}) \geq 0$.

(ii) $(X,\alpha,\mathbb{T})$ is $\mathbb{T}$*–equivariant* $\boldsymbol{F}^{\mathrm{NA}}$*–polystable* (on test configurations of C) if it is $\mathbb{T}$–equivariant $\boldsymbol{F}^{\mathrm{NA}}$–semistable and, further, $\boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}) = 0$ if and only if $(\mathscr{X},\mathscr{A})$ is a product test configuration.

(iii) $(X,\alpha,\mathbb{T})$ is $\mathbb{T}^{\mathbb{C}}$*–uniform* $\boldsymbol{F}^{\mathrm{NA}}$*–stable* (on test configurations of C) if there exists a uniform positive constant $\lambda > 0$ such that, for any test configuration $(\mathscr{X},\mathscr{A}) \in C$,
$$\boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}) \geq \lambda \boldsymbol{J}^{\mathrm{NA}}_{\mathbb{T}^{\mathbb{C}}}(\mathscr{X},\mathscr{A}),$$
where $\boldsymbol{J}^{\mathrm{NA}}_{\mathbb{T}^{\mathbb{C}}}(\mathscr{X},\mathscr{L})$ is as introduced in (20).

Remark 4.8 If $\boldsymbol{F}$ is bounded below, $(X,\alpha,\mathbb{T})$ is $\mathbb{T}$–equivariant $\boldsymbol{F}^{\mathrm{NA}}$–semistable. Furthermore both (ii) and (iii) imply (i) and, in the polarized case, (iii) implies (ii) by the results in [18; 49; 57].

Theorem 4.9 [14; 49; 57; 66] *Suppose $\boldsymbol{F}$ is a functional defined on the space of $\mathbb{T}$–invariant Kähler metrics in α, which is $\mathbb{T}$–relatively $\mathbb{T}^{\mathbb{C}}$–proper. Suppose, furthermore, that $\boldsymbol{F}$ admits a non-Archimedean version $\boldsymbol{F}^{\mathrm{NA}}$ defined for a class C of $\mathbb{T}$–equivariant Kähler test configurations of $(X,\alpha,\mathbb{T})$. Then $(X,\alpha,\mathbb{T})$ is $\mathbb{T}$–equivariant $\boldsymbol{F}^{\mathrm{NA}}$–polystable on C. If, moreover, (X,L) is a polarized variety and $\alpha = 2\pi c_1(L)$, then $(X,\alpha,\mathbb{T})$ is $\mathbb{T}^{\mathbb{C}}$–uniform $\boldsymbol{F}^{\mathrm{NA}}$–stable on polarized test configurations in C.*

Proof For the first part, we follow [66] with some minor modifications. We want to show that if $\boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}) = 0$, then $(\mathscr{X},\mathscr{A})$ is a product test configuration.

Fix a $\mathbb{T}\times\mathbb{S}^1$–invariant Kähler form $\Omega \in \mathscr{A}$ and let ω_t be the corresponding ray of smooth $\mathbb{T}$–invariant Kähler forms in α, and $\psi_t \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ the *normalized* smooth ray of Kähler potentials satisfying $\boldsymbol{I}(\psi_t) = 0$. According to [65], the Kähler test configuration $(\mathscr{X},\mathscr{A})$ also determines a unique $C^{1,\bar{1}}$ weak geodesic ray φ_t in $\mathcal{K}^{1,\bar{1}}(X,\omega_0)$, emanating from ψ_0. Furthermore, φ_t is invariant under $\mathbb{T}$ (by its uniqueness) provided that we have $\psi_0 \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$. According to [66, Proposition 4.2], we can consider instead of $\mathscr{A}$ the *relative Kähler class* $\mathscr{A}_c = \mathscr{A} - c[X_0] = \mathscr{A} - c\pi^*(\mathcal{O}_{\mathbb{P}^1}(1))$ (for a constant c determined from $\mathscr{A}$ and where $[X_0]$ denotes the divisor corresponding to the central fibre X_0 of $\mathscr{X}$) such that the $C^{1,\bar{1}}$ weak geodesic ray φ^c_t corresponding to $(\mathscr{X},\mathscr{A}_c)$ is the projection of φ_t to the slice $\mathcal{K}^{1,\bar{1}}_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$. Notice that the smooth $(1,1)$–form $\Omega - c\pi^*\omega_{\mathrm{FS}} \in \mathscr{A}_c$ defines the same smooth ray ω_t of $\mathbb{T}$–invariant Kähler metrics, and thus the same ray of smooth potentials ψ_t is in $\mathcal{K}_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$ and $\boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}_c) = \boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}) = 0$. The key point is that (17) and
$$\lim_{t\to\infty} \frac{\boldsymbol{F}(\omega_{\psi_t})}{t} = \boldsymbol{F}^{\mathrm{NA}}(\mathscr{X},\mathscr{A}_c) = 0$$
yield an estimate $0 \leq d_1^{[\mathbb{G}]}(0,\psi_t) \leq o(t)$, which is shown in [66, Lemma 4.8] to be equivalent to $0 \leq d_1^{[\mathbb{G}]}(0,\varphi^c_t) \leq o(t)$. We can now apply the arguments in the proof of

the implication "(2) $\Longrightarrow$ (5)" of [66, Theorem 4.4] by replacing the Mabuchi energy with the abstract functional $\boldsymbol{F}$ and the group $\mathrm{Aut}_0(X)$ with $\mathbb{T}^{\mathbb{C}}$, and noting that in our $\mathbb{T}$–relative situation instead of the cscK potential ψ_0 in [66, Proposition 4.10] we can take *any* Kähler potential in $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ (as ω_{ψ_0} is $\mathbb{T}$–invariant and $\mathbb{T}^{\mathbb{C}}$ is reductive). We thus deduce the implication (5) of [66], namely, that the geodesic ray φ_t^c associated to $(\mathscr{X},\mathscr{A}_c)$ is given by the ω_0–relative Kähler potentials of $\exp(tJ\xi)^*(\omega_{\psi_0})$ in $\boldsymbol{I}^{-1}(0)$, where ξ is a vector field in the Lie algebra of $\mathbb{T}$; it follows from [66, Theorem A.6] that $(\mathscr{X},\mathscr{A}_c)$, and hence also $(\mathscr{X},\mathscr{A})$, is a product test configuration.

The second part follows immediately from (18) and Example 4.3 (continued). $\square$

We next apply Theorem 4.9 to $\boldsymbol{F}=\boldsymbol{M}_{v,w}$ and $\boldsymbol{F}^{\mathrm{NA}}=\mathscr{F}_{v,w}$.

Definition 4.10 Let $\boldsymbol{F}^{\mathrm{NA}}=\mathscr{F}_{v,w}$, where $\mathscr{F}_{v,w}$ is defined on any smooth $\mathbb{T}$–equivariant test configuration via the formula (21); see Definition 4.5. We then refer to the $\boldsymbol{F}^{\mathrm{NA}}$–stability notions introduced in Definition 4.7(i)–(iii) as $\mathbb{T}$–equivariant (v,w)–K–semistability, $\mathbb{T}$–equivariant (v,w)–K–polystability, and $\mathbb{T}^{\mathbb{C}}$–uniform (v,w)–K–stability, respectively, on $\mathbb{T}$–invariant dominating smooth Kähler test configurations with reduced central fibre.

Proof of Corollary 1 modulo Theorem 1 By definition of $\boldsymbol{M}_{v,w}$ (see Definition 1.1),

$$\boldsymbol{M}_{v,w}(\varphi+c)=\boldsymbol{M}_{v,w}(\varphi)+c\int_X(\mathrm{Scal}_v(\omega_\varphi)-w(\mu_\varphi))\omega_\varphi^{[m]},$$

showing that if $\boldsymbol{M}_{v,w}$ is bounded below on $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ (in particular if $\boldsymbol{M}_{v,w}$ is $\mathbb{T}$–relatively $\mathbb{T}^{\mathbb{C}}$–proper), then the relation (7) holds and $\boldsymbol{M}_{v,w}$ defines a functional on the space of $\mathbb{T}$–invariant Kähler metrics in α (see Remark 1.2). In this case, Example 4.4 tells us that $\mathscr{F}_{v,w}(\mathscr{X},\mathscr{A})$ defines a non-Archimedean version of $\boldsymbol{M}_{v,w}$. We can now apply Theorem 4.9. $\square$

5 Semisimple principal fibrations

Let (X,ω) be a compact Kähler $2m$–manifold, endowed with a hamiltonian isometric action of an r–dimensional torus $\mathbb{T}$. As $\mathbb{T}$ will act on various spaces, we shall use at times super- and subscripts to emphasize the space on which $\mathbb{T}$ acts. For instance, $\mathbb{T}_X$ will denote the $\mathbb{T}$–action on X. Let $\mathfrak{t}$ be the Lie algebra of $\mathbb{T}$ and $\Lambda\subset\mathfrak{t}$ the lattice of generators of circle groups in $\mathbb{T}$ (ie $\mathbb{T}=\mathfrak{t}/2\pi\Lambda$). We denote by $\mu_\omega\colon X\to\Delta\subset\mathfrak{t}^*$ the normalized $\mathbb{T}_X$–momentum map of ω (the map whose image is a fixed compact convex polytope $\Delta\subset\mathfrak{t}^*$).

Let $B = B_1 \times \cdots \times B_k$ be a $2n$–dimensional cscK manifold where each (B_a, ω_{B_a}), for $a = 1, \dots, k$, is a compact cscK Hodge Kähler $2n_a$–manifold (ie $\frac{1}{2\pi}[\omega_{B_a}]$ is in $H^2(B_a, \mathbb{Z})$), and $\pi_B \colon P \to B$ a principal $\mathbb{T}$–bundle endowed with a connection 1–form $\theta \in \Omega^1(P, \mathfrak{t})$ with curvature

$$d\theta = \sum_{a=1}^{k} (\pi_B^* \omega_{B_a}) \otimes p_a \quad \text{for } p_a \in \Lambda. \tag{22}$$

Remark 5.1 The principal $\mathbb{T}$–bundle P above can be described in terms of r complex line bundles over B as follows. Fixing a lattice basis $\{\xi_1, \dots, \xi_r\}$ of $\mathfrak{t}$ and writing $p_a = \sum_{i=1}^r p_{ai} \xi_i$ for $p_{ai} \in \mathbb{Z}$ with $a = 1, \dots, k$ (22) yields that P is the (fiberwise) product of r principal U(1)–bundles $P_i \to B$, where each P_i is associated to a complex line bundle L_i^* on B with Chern class $2\pi c_1(L_i^*) = -\sum_{a=1}^r p_{ai} \pi_B^*[\omega_{B_a}]$:

$$2\pi c_1(P) := -2\pi \sum_{i=1}^{r} c_1(L_i^*) \otimes \xi_i = \sum_{a=1}^{k} \pi_B^*[\omega_{B_a}] \otimes p_a.$$

Fixing a connection 1–form θ on P as in (22) amounts to introducing a hermitian metric h_i^* on each L_i^*, with curvature $-\sum_{a=1}^r p_{ai} \pi_B^*(\omega_{B_a})$, and identifying $P_i \subset L_i^*$ with the corresponding unitary $\mathbb{S}^1$–bundle.

Let $\mathcal{D} = \mathrm{ann}(\theta) \subset TP$ be the horizontal distribution defined by θ, leading to a splitting

$$TP = \mathcal{D} \oplus \mathfrak{t}_P,$$

where $\mathfrak{t}_P$ denotes the Lie algebra of $\mathbb{T}_P$ inside $C^\infty(P, TP)$, corresponding to the $\mathbb{T}$–action $\mathbb{T}_P$ on P. The lift J_B of the integrable almost complex structure of B to $\mathcal{D}$ gives rise to a CR structure $(\mathcal{D}, J_B)$ on P (of codimension r).

We further let $Z := X \times P$ and consider the induced $\mathbb{T}$–action, denoted by $\mathbb{T}_Z$, generated by $(-\xi_i^X + \xi_i^P)$ for any basis of Λ as above. We thus define

$$Y := Z/\mathbb{T}_Z.$$

It follows that Y is a $2(m+n)$–dimensional smooth manifold, and $\pi_Y \colon Z = X \times P \to Y$ is a principal $\mathbb{T}$–bundle over Y, whereas $\pi_B \colon P \to B$ defines a fibration $\pi_B \colon Y \to B$ with smooth fibres X, as summarized in the diagram

$$\begin{tikzcd} & Z = X \times P \arrow[dl, "/\mathbb{T}_P"'] \arrow[dr, "/\mathbb{T}_Z"] \arrow[dd, "\pi_B"] & \\ X \times B \arrow[dr, "\pi_B"'] & & Y \arrow[dl, "\pi_B"] \\ & B & \end{tikzcd}$$

The $\mathbb{T}_X$–action on the factor X in $Z = X \times P$ descends to a $\mathbb{T}$–action on Y, denoted by $\mathbb{T}_Y$, which preserves each fibre (and thus coincides with the action of $\mathbb{T}_X$). Notice that the 1–form θ also defines a connection 1–form on Z with horizontal distribution $\mathscr{H}$:

$$T(X \times P) = \mathscr{H} \oplus \mathfrak{t}_Z, \quad \mathscr{H} = TX \oplus \mathcal{D} = \operatorname{ann}(\theta). \tag{23}$$

This gives rise to an induced CR structure $(\mathscr{H}, J = J_X \oplus J_B)$ of codimension r on Z, which is clearly invariant under the $\mathbb{T}_Z$–action, and therefore defines a $\mathbb{T}_Y$–invariant complex structure J_Y on Y.

We now consider Kähler metrics on Y, compatible with the fibre bundle construction of the above form. To simplify the notation, we denote by $\omega_a := \omega_{B_a}$ the (fixed) cscK metric on each factor B_a, by ω a $\mathbb{T}$–invariant Kähler structure in the class α on X, and by $\widetilde{\omega}$ the resulting Kähler structure on Y, which is defined in terms of a basic 2–form on $Z = X \times P$, depending on k real constants $c_a \in \mathbb{R}$ (which will be fixed) such that, for each $a = 1, \dots, k$, the affine-linear function $\langle p_a, \mu \rangle + c_a$ on $\mathfrak{t}^*$ is strictly positive on the momentum image Δ:

$$\begin{aligned} \widetilde{\omega} &:= \omega + \sum_{a=1}^{k} (\langle p_a, \mu_\omega \rangle + c_a) \pi_B^* \omega_a + \langle d\mu_\omega \wedge \theta \rangle \\ &= \omega + \sum_{a=1}^{k} c_a (\pi_B^* \omega_a) + d(\langle \mu_\omega, \theta \rangle). \end{aligned} \tag{24}$$

In the above expression, $\langle \cdot, \cdot \rangle$ stands for the natural pairing between $\mathfrak{t}$ and $\mathfrak{t}^*$. Thus $\langle p_a, \mu_\omega \rangle$ is a smooth function, $\langle \mu_\omega, \theta \rangle$ is a 1–form, and $\langle d\mu_\omega \wedge \theta \rangle$ is a 2–form on Z. One can directly check from the above expression that $\widetilde{\omega}$ is closed and $\mathbb{T}_Z$–basic, and is positive definite on $(\mathscr{H}, J_X \oplus J_B)$, so it is the pullback of a Kähler form on Y. We shall tacitly identify in the sequel the Kähler form on Y with its pullback (24) on $Z = X \times P$. Notice that $\widetilde{\omega}$ is $\mathbb{T}_Y$–invariant and μ_ω, seen as a smooth $\mathbb{T}_Z$–invariant function on Z, is the Δ–normalized momentum map.

Remark 5.2 The horizontal part $\widetilde{\omega}_h := \widetilde{\omega}|_{\mathscr{H}}$ of the 2–form $\widetilde{\omega}$ on $Z = X \times P \xrightarrow{\pi_B} X \times B$ is invariant and basic with respect to the action $\mathbb{T}_P$ on the factor P, and thus induces a hermitian (non-Kähler in general) metric on $X \times B = X \times \prod_{a=1}^{k} B_a$, given by

$$\tilde{\omega}_h = \omega + \sum_{a=1}^{k} (\langle p_a, \mu_\omega \rangle + c_a) \omega_a,$$

which is an instance of warped geometry. On can thus think of $(X \times B, \widetilde{\omega}_h)$ and $(Y, \widetilde{\omega})$ as being related by the twist construction of [69] applied to $(Z, \widetilde{\omega}, \mathbb{T}_Z)$ and $(Z, \widetilde{\omega}, \mathbb{T}_P)$.

Definition 5.3 The Kähler manifold $(Y, \mathbb{T}_Y)$ constructed as above will be called a *semisimple* $(X, \mathbb{T})$*–principal fibration* associated to the Kähler manifold $(X, \mathbb{T})$ and the product cscK manifold $B = B_1 \times \cdots \times B_k$. The $\mathbb{T}_Y$–invariant Kähler metric $\tilde{\omega}$ on Y constructed from a $\mathbb{T}_X$–invariant Kähler metric ω on X (and fixed cscK metrics ω_a on B_a) will be called *bundle-compatible*.

Remark 5.4 In the case when $(X, \mathbb{T}, \omega)$ is a toric Kähler manifold, a semisimple $(X, \mathbb{T})$–principal fibration endowed with a bundle-compatible Kähler metric is an example of a semisimple rigid toric fibration in the sense of [6], and is thus described by the *generalized Calabi construction* with a global product structure on the base and no blow-downs.

5.1 The space of functions

The above bundle construction gives rise to a natural embedding of the space $C^\infty_{\mathbb{T}}(X)$ of $\mathbb{T}_X$–invariant smooth functions on X to the space $C^\infty_{\mathbb{T}}(Y)$ of $\mathbb{T}_Y$–invariant smooth functions on Y: for any $\varphi \in C^\infty_{\mathbb{T}}(X)$ we consider the induced function on $Z = X \times P$, which is clearly $\mathbb{T}_Z$–invariant, and thus descends to a smooth $\mathbb{T}_Y$–invariant function on Y. We shall tacitly identify φ and its image in $C^\infty_{\mathbb{T}}(Y)$, ie we shall consider

$$C^\infty_{\mathbb{T}}(X) \subset C^\infty_{\mathbb{T}}(Y).$$

Notice that the above embedding is closed in the Fréchet topology, as we can identify a smooth $\mathbb{T}_X$–invariant function on X with a smooth $\mathbb{T}_Y$–invariant function φ on Y, which has the property

$$d_P(\pi_Y^* \varphi) = 0$$

on $Z = X \times P$.

More generally, for any $\mathbb{T}_Y$–invariant smooth function $\psi \in C^\infty_{\mathbb{T}}(Y)$, its lift $\pi_Y^* \psi$ to $Z = X \times P$ is a smooth function which is both $\mathbb{T}_Z$– and $\mathbb{T}_X$–invariant, or equivalently $\mathbb{T}_X$– and $\mathbb{T}_P$–invariant. It thus follows that $\pi_Y^* \psi$ can be equivalently viewed as a $\mathbb{T}_X$–invariant smooth function on $X \times B$, ie we have an identification

$$C^\infty_{\mathbb{T}}(Y) \cong C^\infty_{\mathbb{T}}(X \times B). \tag{25}$$

In particular, for any fixed point $x \in X$, we shall denote by $\psi_x \in C^\infty(B)$ the induced smooth function on B, and for any fixed point $b \in B$ by $\psi_b \in C^\infty_{\mathbb{T}}(X)$ the induced function on X. We thus have the identification

$$C^\infty_{\mathbb{T}}(X) \cong \{\psi \in C^\infty_{\mathbb{T}}(Y) \mid d_B \psi_x = 0 \text{ for all } x \in X\}.$$

5.2 The space of bundle-compatible Kähler metrics

We shall next use the construction of (24) in order to identify the space $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ of $\mathbb{T}_X$–invariant ω_0–relative Kähler potentials on X as a subset of the space $\mathcal{K}_{\mathbb{T}}(Y,\tilde{\omega}_0)$ of $\mathbb{T}_Y$–invariant $\tilde{\omega}_0$–relative Kähler potentials on Y.

Lemma 5.5 *Let $\omega_\varphi=\omega_0+d_Xd_X^c\varphi$ be a $\mathbb{T}_X$–invariant Kähler form on X in the Kähler class $\alpha=[\omega_0]$, where $\varphi\in\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ is a $\mathbb{T}_X$–invariant smooth function on X. Denote by μ_φ the momentum map of $\mathbb{T}_X$ with respect to ω_φ, satisfying the normalization $\mu_\varphi(X)=\Delta$, and by $\tilde{\omega}_\varphi$ the induced Kähler metric on Y, via* (24). *Then*

$$\tilde{\omega}_\varphi=\tilde{\omega}_0+d_Yd_Y^c\varphi,$$

where φ stands for the induced smooth function on Y.

Proof Recall that $\mu_\varphi=\mu_0+d_X^c\varphi$; see (5). By (24), the pullback of $\tilde{\omega}_\varphi$ to $Z=X\times P$ is

$$\begin{aligned}\tilde{\omega}_\varphi&=\omega_\varphi+\sum_{a=1}^k c_a(\pi_B^*\omega_a)+d_Z\langle\mu_\varphi,\theta\rangle=\omega_0+\sum_{a=1}^k c_a(\pi_B^*\omega_a)+d_Xd_X^c\varphi+d_Z\langle\mu_\varphi,\theta\rangle\\&=\tilde{\omega}_0+d_Zd_X^c\varphi+d_Z(\langle d_X^c\varphi,\theta\rangle),\end{aligned}$$

so it is enough to check that

$$d_Y^c\varphi=d_X^c\varphi+\langle d_X^c\varphi,\theta\rangle \tag{26}$$

for any $\mathbb{T}_X$–invariant smooth function φ on X. To this end, let us choose a basis $\{\xi_1,\dots,\xi_r\}$ of $\mathfrak{t}$, with dual basis $\{\xi^1,\dots,\xi^r\}$ of $\mathfrak{t}^*$, and write

$$d_X^c\varphi=\sum_{j=1}^r(d_X^c\varphi)(\xi_j^X)\xi^j\quad\text{and}\quad\theta=\sum_{j=1}^r\theta_j\xi_j$$

for 1–forms θ_j on Z such that θ_j is zero on $\mathscr{H}$ and $\theta_j(\xi_i^P)=\theta_j(-\xi_i^X+\xi_i^P)=\delta_{ij}$. Thus, (26) is equivalent to

$$d_Y^c\varphi=d_X^c\varphi+\sum_{j=1}^r(d_X^c\varphi)(\xi_j^X)\theta_j.$$

Evaluating the right side of the above equality on the generators $(-\xi_j^X+\xi_j^P)$ of $\mathfrak{t}_Z$, we see that it is a π_Y–basic 1–form on Z, and thus is the pullback of a 1–form on Y via π_Y. The claim follows easily. □

Thus, Lemma 5.5 defines an embedding $\mathcal{K}_{\mathbb{T}}(X,\omega_0)\subset\mathcal{K}_{\mathbb{T}}(Y,\tilde{\omega}_0)$ and we have also identified in Section 5.1 a natural embedding of the space of $\mathbb{T}_X$–invariant functions on X into the space of $\mathbb{T}_Y$–invariant functions on Y, through their pullbacks to $Z=X\times P$.

Letting $\theta := \sum_{j=1}^{r} \theta_j \otimes \xi_j^P$ be the decomposition of the connection 1–form θ on P in a basis $\{\xi_1, \dots, \xi_r\}$ of the lattice $\Lambda \subset \mathfrak{t}$, and $\theta^{\wedge r} := \theta_1 \wedge \cdots \wedge \theta_r$, it follows from (24) and Lemma 5.5 that for any $\varphi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{T}}(Y, \tilde{\omega}_0)$, the measure $\tilde{\omega}_\varphi^{[m+n]}$ on Y is the pushforward of the measure on Z

$$\frac{1}{(2\pi)^r} \tilde{\omega}_\varphi^{[m+n]} \wedge \theta^{\wedge r} = \frac{1}{(2\pi)^r} \big(p(\mu_\varphi) \omega_\varphi^{[m]} \wedge \textstyle\bigwedge_{a=1}^{k} \pi_B^* \omega_a^{[n_a]} \big) \wedge \theta^{\wedge r}, \tag{27}$$

where

$$p(\mu) := \prod_{a=1}^{k} (\langle p_a, \mu \rangle + c_a)^{n_a} \quad \text{for } n_a = \dim_{\mathbb{C}}(B_a) \tag{28}$$

is a positive polynomial on Δ, determined by the semisimple $(X, \mathbb{T})$–principal fibration Y and the given bundle-compatible Kähler class on it. It thus follows that any $\mathbb{T}_X$–invariant integrable function f on X defines an integrable $\mathbb{T}_Y$–invariant function on Y and, for any $\varphi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{T}}(Y, \tilde{\omega}_0)$,

$$\int_Y f \tilde{\omega}_\varphi^{[n+m]} = \mathrm{Vol}(B, \omega_B) \int_X p(\mu_\varphi) f \omega_\varphi^{[m]}. \tag{29}$$

Corollary 5.6 *There exists an embedding $\mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{T}}(Y, \tilde{\omega}_0)$ such that, for any smooth curve $\psi_t \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{T}}(Y, \tilde{\omega}_0)$,*

$$L_1^Y(\psi_t) = \mathrm{Vol}(B, \omega_B) L_{1,p}^X(\psi_t),$$

where $p(\mu)$ is the positive weight function on Δ defined in (28), *$L_{1,p}^X$ is the $p(\mu)$–weighted length function on $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$ given by*

$$L_{1,p}^X(\psi_t) := \int_0^1 \Big(\int_X |\dot{\psi}_t| p(\mu_{\psi_t}) \omega_{\psi_t}^{[m]} \Big) \, dt,$$

and L_1^Y is the length function on $\mathcal{K}_{\mathbb{T}}(Y, \tilde{\omega}_0)$ corresponding to the weight $p = 1$. In particular, $d_1^Y(\varphi_0, \varphi_1) = \mathrm{Vol}(B, \omega_B) d_{1,p}^X(\varphi_0, \varphi_1)$ for any $\varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{T}}(Y, \tilde{\omega}_0)$, where $d_{1,p}^X$ is the induced distance via the length functional $L_{1,p}^X$.

Proof This is a direct consequence of (29). □

Lemma 5.7 *Let φ be a smooth $\mathbb{T}_X$–invariant function on X, also considered as a smooth $\mathbb{T}_Y$–invariant function on Y, and ω be a $\mathbb{T}_X$–invariant Kähler metric on X with $\tilde{\omega}$ the corresponding $\mathbb{T}_Y$–invariant Kähler metric on Y given by* (24). *Then*

$$\|d\varphi\|_\omega^2 = \|d\varphi\|_{\tilde{\omega}}^2.$$

Proof We use that

$$\|d\varphi\|^2_{\omega} = \frac{d_X\varphi \wedge d^c_X\varphi \wedge \omega^{[m-1]}}{\omega^{[m]}} = \frac{d_X\varphi \wedge d^c_X\varphi \wedge \omega^{[m-1]} \wedge p(\mu_\omega)(\pi^*_B\omega_B)^{[n]} \wedge \theta^{\wedge r}}{\omega^{[m]} \wedge p(\mu_\omega)(\pi^*_B\omega_B)^{[n]} \wedge \theta^{\wedge r}},$$

$$\|d\varphi\|^2_{\tilde\omega} = \frac{d_Y\varphi \wedge d^c_Y\varphi \wedge \tilde\omega^{[m+n-1]}}{\tilde\omega^{[m+n]}} = \frac{d_Y\varphi \wedge d^c_Y\varphi \wedge \tilde\omega^{[m+n-1]} \wedge \theta^{\wedge r}}{\tilde\omega^{[m+n]} \wedge \theta^{\wedge r}}$$

(where the right sides are written on $X \times P$) together with $d_X\varphi = d_Y\varphi$ and (27). □

Proposition 5.8 *The embedding in Corollary 5.6 is totally geodesic with respect to the weak $C^{1,\bar 1}$ geodesics.*

Proof Let $\varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$. If φ_0 and φ_1 can be connected by a *smooth* geodesic φ_t, ie with a smooth curve in $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$ such that

$$\ddot\varphi = \|d\dot\varphi\|^2_{\omega_\varphi}, \tag{30}$$

then, by Lemma 5.7, φ_t is also a smooth geodesic in $\mathcal{K}_{\mathbb{T}}(Y, \tilde\omega_0)$ connecting φ_0 and φ_1 in $\mathcal{K}_{\mathbb{T}}(Y, \tilde\omega_0)$.

In general, by the results in [22], φ_0 and φ_1 can be connected only with a weak $C^{1,\bar 1}$–geodesic in $\mathcal{K}^{1,\bar 1}_{\mathbb{T}}(X, \omega_0)$, where $\mathcal{K}^{1,\bar 1}(X, \omega_0)$ stand for the space of $C^1(X)$ functions φ on X such that $\omega_0 + dd^c\varphi \ge 0$ and has bounded coefficients as a $(1,1)$–current. More precisely, letting $\Sigma := \{1 < z < e\} \subset \mathbb{C}$, it is shown in [22] that there exists a unique weak solution (ie a positive $(1,1)$–current in the sense of Bedford and Taylor) of the homogeneous Monge–Ampère equation

$$\begin{gathered}(\pi^*_X\omega_0 + d_Xd^c_X\Phi)^{m+1} = 0, \quad \pi^*_X\omega_0 + d_Xd^c_X\Phi \ge 0 \quad \text{for } \Phi \in C^{1,\alpha}(X \times \bar\Sigma), \\ \Phi(x,1) = \varphi_0(x), \quad \Phi(x,e) = \varphi_1(x).\end{gathered} \tag{31}$$

It was later shown in [25] that Φ is actually of regularity $C^{1,1}(X \times \bar\Sigma)$. Note that, by uniqueness, Φ is $\mathbb{T}$–invariant as soon as φ_0 and φ_1 are. The link with (30) is (see [64]) that if Φ were actually smooth, we could recover the smooth geodesic φ_t joining φ_0 and φ_1 by letting $t := \log|z|$ and $\varphi_t(x) := \Phi(x, \log|z|)$. In the general case, the curve φ_t of (weak) ω_0–relative plurisubharmonic potentials (of regularity $C^{1,1}(X \times [0,1])$) is referred to as the *weak* $C^{1,\bar 1}$–geodesic joining φ_0 and φ_1.

We are thus going to check that any weak $C^{1,\bar 1}$–geodesic on X (invariant under $\mathbb{T}_X$) defines, via Lemma 5.5, a $C^{1,\bar 1}$–geodesic on Y. To this end, we need to show that Φ satisfies

$$(\pi^*_Y\tilde\omega_0 + d_Yd^c_Y\Phi)^{m+n+1} = 0 \quad \text{and} \quad \pi^*_Y\tilde\omega_0 + d_Yd^c_Y\Phi \ge 0, \tag{32}$$

the regularity statements being automatically satisfied on Y.

By the results in [16; 22], Φ can be approximated as $\varepsilon \to 0$, both in the weak sense of currents and in $C^{1,\alpha}(X \times \overline{\Sigma})$ (for a fixed $\alpha \in (0,1)$), by smooth functions $\Psi^{\varepsilon}(x,z)$ on $X \times \overline{\Sigma}$ for $\varepsilon > 0$ which solve

$$
\begin{gathered}
(\pi_X^* \omega_0 + d_X d_X^c \Psi^{\varepsilon})^{[m+1]} = \varepsilon((\pi_X^* \omega_0)^{[m]} \wedge (dx \wedge dy)), \\
\pi_X^* \omega_0 + d_X d_X^c \Psi^{\varepsilon} > 0, \quad \Psi^{\varepsilon}(x,1) = \varphi_0, \quad \Psi^{\varepsilon}(x,e) = \varphi_1.
\end{gathered}
\tag{33}
$$

By the uniqueness of the smooth solution of (33) (and using that both φ_0 and φ_1 are $\mathbb{T}_X$–invariant), we have that $\Psi^{\varepsilon}(x,z)$ is a $\mathbb{T}_X$–invariant smooth function on X for any $z \in \overline{\Sigma}$. Furthermore, the positivity condition on the second line yields that $\Psi^{\varepsilon}(x,z) \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ for any $z \in \overline{\Sigma}$. We can then also see $\Psi^{\varepsilon}(x,z)$, via its pullback to $X \times P \times \overline{\Sigma}$, as a $\mathbb{T}_Y$–invariant smooth function on $Y \times \overline{\Sigma}$; the arguments in the proof of Lemma 5.5 yield that $\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon} > 0$ on $Y \times \overline{\Sigma}$. Also, by the same proof, we have the equality of volume forms, on $X \times P \times \overline{\Sigma}$,

$$
\begin{aligned}
(\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon})^{[m+n+1]} \wedge \theta^{\wedge r} & \\
&= p(\mu_{\Psi^{\varepsilon}})(\pi_X^* \omega_0 + d d^c \Psi^{\varepsilon})^{[m+1]} \wedge (\pi_B^* \omega_B)^{[n]} \wedge \theta^{\wedge r} \\
&= \varepsilon p(\mu_{\Psi^{\varepsilon}})(\pi_X^* \omega_0)^{[m+1]} \wedge (\pi_B^* \omega_B)^{[n]} \wedge \theta^{\wedge r},
\end{aligned}
\tag{34}
$$

where, we recall, $p(\mu) := \prod_{a=1}^{k}(\langle p_a, \mu \rangle + c_a)^{n_a}$ and $\theta^{\wedge r} := \theta_1 \wedge \cdots \wedge \theta_r$ (for $\theta = \sum_{i=1}^{r} \theta_i \otimes \xi_i^P$ with respect to a basis $\{\xi_1, \ldots, \xi_r\}$ of $\Lambda \subset \mathfrak{t}$), and, for any fixed $z \in \overline{\Sigma}$, $\mu_{\Psi^{\varepsilon}}$ denotes the normalized $\mathbb{T}_X$–momentum map (5) of $\omega_0 + d_X d_X^c \Psi^{\varepsilon}$. Notice that, as p is uniformly bounded on Δ by positive constants, it follows by (34) that

$$
\lim_{\varepsilon \to 0}((\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon})^{[m+n+1]} \wedge \theta^{\wedge r}) = 0
$$

weakly (as measures on $Z \times \overline{\Sigma}$). The measure $(\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon})^{[m+n+1]}$ is the pushforward measure of $(\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon})^{[m+n+1]} \wedge \theta^{\wedge r}$ to Y, so we obtain, on Y,

$$
\lim_{\varepsilon \to 0}((\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon})^{[m+n+1]}) = 0.
$$

Furthermore, using the $C^{1,\alpha}$–convergence of Ψ^{ε} to Φ, we get the weak convergences (of positive $(1,1)$–currents)

$$
\begin{gathered}
\lim_{\varepsilon \to 0}(\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon}) = \pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Phi \geq 0, \\
0 = \lim_{\varepsilon \to 0}(\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Psi^{\varepsilon})^{[m+n+1]} = (\pi_Y^* \tilde{\omega}_0 + d_Y d_Y^c \Phi)^{[m+n+1]}.
\end{gathered}
$$

Thus, (32) follows. ∎

Lemma 5.9 *Let v be a smooth positive weight function on Δ, let ω and $\tilde{\omega}$ be $\mathbb{T}$–invariant Kähler metrics on X and Y, respectively, given by (24), and suppose (B_a, ω_a) has constant scalar curvature $\mathrm{Scal}(\omega_a) = s_a$. Then the v–scalar curvature $\mathrm{Scal}(\tilde{\omega})$, considered as smooth function on $X \times P$, is given by*

$$\mathrm{Scal}_v(\tilde{\omega}) = \frac{1}{p(\mu_\omega)} \mathrm{Scal}_{pv}(\omega) + v(\mu_\omega) q(\mu_\omega), \tag{35}$$

with $p(\mu) = \prod_{a=1}^{k} (\langle p_a, \mu\rangle + c_a)^{n_a}$ and $q(\mu) = \sum_{a=1}^{k} s_a / (\langle p_a, \mu\rangle + c_a)$. In particular, ω is a $(pv, \tilde{w})$–cscK metric on X if and only if $\tilde{\omega}$ is a (v, w)–cscK metric on Y, with

$$\tilde{w}(\mu) = p(\mu)(w(\mu) - v(\mu) q(\mu)).$$

Proof We apply the arguments in the proof of [3, Proposition 7] to both $(X, \mathbb{T}_X)$ and $(Y, \mathbb{T}_Y)$ to compute the corresponding scalar curvatures, and compare the results.

On X, we consider the open dense subset $\mathring{X} \subset X$ of stable points of the $\mathbb{T}_X$–action, and take the quotient $\pi_S \colon \mathring{X} \to S := \mathring{X} / \mathbb{T}_X^{\mathbb{C}}$ under the induced complexified action $\mathbb{T}_X^{\mathbb{C}} \cong (\mathbb{C}^*)^r$ (thus S is a complex $2(m-r)$–dimensional orbifold).

Consider the pointwise ω–orthogonal and $\mathbb{T}$–invariant decomposition

$$T\mathring{X} = \mathcal{H} \oplus \mathfrak{t}_X \oplus J\mathfrak{t}_X,$$

and write the Kähler structure (g, J, ω) on X as

$$g = g_{\mathcal{H}} + \sum_{i,j=1}^{r} H_{ij} (\eta_i \otimes \eta_j + J\eta_i \otimes J\eta_j), \quad \omega = \omega_{\mathcal{H}} + \sum_{i,j=1}^{r} H_{ij} \eta_i \wedge J\eta_j,$$

where, for a fixed basis $\{\xi_1, \dots, \xi_r\}$ of $\mathfrak{t}$, the 1–forms η_j on $\mathring{X}$ are defined by $(\eta_j)|_{\mathcal{H}} = 0$, $\eta_j(\xi_i^X) = \delta_{ij}$; $\eta_j(J\xi_i^X) = 0$ and $H_{ij} = g(\xi_i^X, \xi_j^X)$.

We next fix a local volume form Vol_S on S in some holomorphic coordinates, and pointwise write

$$\omega_{\mathcal{H}}^{[m-r]} = Q \pi_S^*(\mathrm{Vol}_S) \tag{36}$$

for some positive (locally defined) smooth function Q on $\mathring{X}$ (where both $\omega_H^{[m-1]}$ and $\pi_S^*(\mathrm{Vol}_S)$ are seen as sections of $\bigwedge^{m-1} \mathcal{H}^*$). According to [3, Proposition 7],

$$\kappa := -\tfrac{1}{2}(\log Q + \log\det(H_{ij})) \tag{37}$$

is a (local) Ricci potential of ω, ie $\rho_\omega = d_X d_X^c \kappa$, and thus

$$\mathrm{Scal}(\omega) = -2 \frac{d_X d_X^c \kappa \wedge \omega^{[m-1]}}{\omega^{[m]}}.$$

We can now make a similar argument on Y, noting that the Kähler reduction of $\mathring{Y}$ by the induced $\mathbb{T}_Y$–action is $S \times B$; taking a local volume form in holomorphic coordinates on $S \times B$ of the form $\mathrm{Vol}_S \wedge \mathrm{Vol}_{B_1} \wedge \cdots \wedge \mathrm{Vol}_{B_k}$, and using (24), we see that a Ricci potential on Y (when pulled back to $X \times P$) is written as

$$\tilde{\kappa} = \sum_{a=1}^{k} \kappa_a - \tfrac{1}{2}(\log \widetilde{Q} + \log\det(H_{ij})),$$

where $\kappa_a := -\frac{1}{2}\log(\omega_a^{[n_a]}/\mathrm{Vol}_{B_a})$ is a Ricci potential of (B_a, ω_a) and

$$\widetilde{Q} = p(\mu_\omega) Q.$$

Thus,

$$\tilde{\kappa} = \sum_{a=1}^{k} \kappa_a + \kappa - \tfrac{1}{2}\log p(\mu_\omega), \tag{38}$$

as functions on $X \times P$. Introducing a basis $(\xi_i)_i$ of Λ and writing the connection 1–form $\theta \in \Omega^1(P, \mathfrak{t})$ as $\theta = \sum_{j=1}^{r} \theta_j \otimes \xi_j^P$ (where the 1–forms θ_j on P are such that θ_j is zero on $\mathcal{D}$ and $\theta_j(\xi_i^P) = \delta_{ij}$), we compute the scalar curvature of $\tilde{\omega}$ to be

$$\mathrm{Scal}(\widetilde{\omega}) = \begin{cases} -(d_Y d_Y^c \tilde{\kappa} \wedge \widetilde{\omega}^{[m+n-1]})/(\widetilde{\omega}^{[m+n]}) & \text{on } Y, \\ -(d_Y d_Y^c \tilde{\kappa} \wedge \widetilde{\omega}^{[m+n-1]} \wedge \theta^{\wedge r})/(\widetilde{\omega}^{[m+n]} \wedge \theta^{\wedge r}) & \text{on } X \times P. \end{cases} \tag{39}$$

By (26) and (38), the pullback of $d_Y d_Y^c \tilde{\kappa}$ to $X \times P$ is given by

$$\begin{aligned} d_Y d_Y^c \tilde{\kappa} &= d_Y d_Y^c\big(\kappa - \tfrac{1}{2}\log p(\mu_\omega)\big) + \sum_{a=1}^{k} d_Y d_Y^c \kappa_a \\ &= d_X d_X^c\big(\kappa - \tfrac{1}{2}\log p(\mu_\omega)\big) + \sum_{j=1}^{r} d_X\big(d_X^c\big(\kappa - \tfrac{1}{2}\log p(\mu_\omega)\big)(\xi_j^X)\big) \wedge \theta_j \\ &\qquad + \sum_{j=1}^{r} d_X^c\big(\kappa - \tfrac{1}{2}\log p(\mu_\omega)\big)(\xi_j) d_P\theta_j + \sum_{a=1}^{k} d d_{B_a}^c \kappa_a \\ &= d_X d_X^c \kappa - \tfrac{1}{2} d_X d_X^c(\log p(\mu_\omega)) \\ &\qquad + \sum_{j=1}^{r} d_X\big(d_X^c\big(\kappa - \tfrac{1}{2}\log p(\mu_\omega)\big)(\xi_j^X)\big) \wedge \theta_j \\ &\qquad + \sum_{a=1}^{k} d_X^c\big(\kappa - \tfrac{1}{2}\log p(\mu_\omega)\big)(p_a)(\pi_B^*\omega_a) + \sum_{a=1}^{k} d d_{B_a}^c \kappa_a, \end{aligned} \tag{40}$$

where in the last equality we used (22) and we have denoted by p_a the induced vector field on X by the element $p_a \in \mathfrak{t}$. We shall compute the term $d_X^c \kappa(p_a)$ on $\mathring{X}$. Using (37), we get

$$d^c_X\kappa(p_a)=\frac{1}{2}\biggl(\frac{\mathcal{L}_{Jp_a}Q}{Q}+\operatorname{tr}(H_{ij}^{-1}(\mathcal{L}_{Jp_a}H_{ij}))\biggr). \tag{41}$$

Taking the wedge product of both sides of (36) with

$$\biggl(\sum_{i,j=1}^{r}H_{ij}\eta_i\wedge J\eta_j\biggr)^{[r]}=\det(H_{ij})\textstyle\bigwedge_{j=1}^{r}(\eta_j\wedge J\eta_j)$$

gives

$$\omega^{[m]}=Q\pi_S^*\operatorname{Vol}_S\wedge\det(H_{ij})\textstyle\bigwedge_{j=1}^{r}(\eta_j\wedge J\eta_j).$$

Applying the Lie derivative $\mathcal{L}_{Jp_a}$ to the above equality yields

$$\begin{aligned}(\Delta_\omega\mu_\omega^{p_a})\omega^{[m]}=(\mathcal{L}_{J\xi_a}Q)\pi_S^*\operatorname{Vol}_S\wedge\det(H_{ij})\textstyle\bigwedge_{j=1}^{r}(\eta_j\wedge J\eta_j)&\\ +Q\pi_S^*\operatorname{Vol}_S\wedge\mathcal{L}_{Jp_a}\bigl(\det(H_{ij})\textstyle\bigwedge_{j=1}^{r}\eta_j\wedge J\eta_j\bigr)&\end{aligned}$$

and

$$\begin{aligned}Q\pi_S^*\operatorname{Vol}_S\wedge\mathcal{L}_{Jp_a}\bigl(\det(H_{ij})\textstyle\bigwedge_{j=1}^{r}(\eta_j\wedge J\eta_j)\bigr)&\\ =\bigl(\operatorname{tr}(H_{ij}^{-1}(\mathcal{L}_{Jp_a}H_{ij}))\bigr)Q\pi_S^*\operatorname{Vol}_S\wedge(\det(H_{ij}))\textstyle\bigwedge_{j=1}^{r}\eta_j\wedge J\eta_j,&\end{aligned}$$

where we used that $\mathcal{L}_{Jp_a}\eta_j$ is a basic form (since $(\mathcal{L}_{Jp_a}\eta_j)(\xi_i)=-\eta_j([Jp_a,\xi_i])=0$). We thus get $\Delta_\omega\mu_\omega^{p_a}=\mathcal{L}_{Jp_a}Q/Q+\operatorname{tr}(H_{ij}^{-1}(\mathcal{L}_{Jp_a}H_{ij}))$ or equivalently, in terms of (41),

$$d^c_X\kappa(p_a)=\tfrac{1}{2}(\Delta_\omega\mu_\omega^{p_a}). \tag{42}$$

Using the above equation in (40), we continue the computation:

$$\begin{aligned}d_Yd^c_Y\tilde\kappa=d_Xd^c_X\kappa-\tfrac{1}{2}dd^c_X(\log p(\mu_\omega))+\sum_{j=1}^{r}d_X\bigl(d^c_X\bigl(\kappa-\tfrac{1}{2}\log p(\mu_\omega)\bigr)(\xi_j^X)\bigr)\wedge\theta_j&\\ +\frac{1}{2}\sum_{a=1}^{k}\biggl(\Delta_\omega\mu_\omega^{p_a}+\frac{\mathcal{L}_{Jp_a}(p(\mu_\omega))}{p(\mu_\omega)}\biggr)(\pi_B^*\omega_a)+\sum_{a=1}^{k}d_{B_a}d^c_{B_a}\kappa_a.&\end{aligned} \tag{43}$$

Recall that, by (27), on Z we have $\tilde\omega^{[m+n]}\wedge\theta^{\wedge r}=p(\mu_\omega)\omega^{[m]}\wedge\bigwedge_{a=1}^{k}\pi_B^*\omega_a^{[n_a]}\wedge\theta^{\wedge r}$. Similarly, by (24),

$$\begin{aligned}&\tilde\omega^{[m+n-1]}\wedge\theta^{\wedge r}\\ &=\sum_{b=1}^{k}\biggl(\frac{p(\mu_\omega)}{\langle\mu_\omega,p_b\rangle+c_b}\omega^{[m]}\wedge(\pi_B^*\omega_b)^{[n_b-1]}\wedge\textstyle\bigwedge_{a=1,a\neq b}^{k}(\pi_B^*\omega_a)^{[n_a]}\wedge\theta^{\wedge r}\biggr)\\ &\qquad+p(\mu_\omega)\omega^{[m-1]}\wedge\textstyle\bigwedge_{a=1}^{k}(\pi_B^*\omega_a)^{[n_a]}\wedge\theta^{\wedge r}.\end{aligned} \tag{44}$$

Using (39), (43), (27) and (44), we obtain

$$
\begin{aligned}
(45)\quad \mathrm{Scal}(\tilde{\omega})
&= \mathrm{Scal}(\omega)+\Delta_\omega(\log p(\mu_\omega)) \\
&\qquad +\sum_{a=1}^{k}\left(\frac{n_a}{\langle\mu_\omega,p_a\rangle+c_a}\left[\Delta_\omega\mu_\omega^{p_a}+\frac{\mathcal{L}_{Jp_a}p(\mu_\omega)}{p(\mu_\omega)}\right]+\frac{s_a}{\langle\mu_\omega,p_a\rangle+c_a}\right) \\
&= \mathrm{Scal}(\omega)+\sum_{a=1}^{k} n_a\Delta_\omega(\log(\langle\mu_\omega,p_a\rangle+c_a)) \\
&\qquad +\sum_{a=1}^{k}\left(\frac{n_a}{\langle\mu_\omega,p_a\rangle+c_a}\left[\Delta_\omega(\langle\mu_\omega,p_a\rangle)+\frac{\mathcal{L}_{Jp_a}(p(\mu_\omega))}{p(\mu_\omega)}\right]+\frac{s_a}{\langle\mu_\omega,p_a\rangle+c_a}\right) \\
&= \mathrm{Scal}(\omega)-\sum_{a,b=1}^{k}\frac{n_a n_b g(p_a,p_b)}{(\langle\mu_\omega,p_a\rangle+c_a)(\langle\mu_\omega,p_b\rangle+c_b)} \\
&\qquad +\sum_{a=1}^{k}\left(\frac{2n_a\Delta_\omega(\langle\mu_\omega,p_a\rangle)}{\langle\mu_\omega,p_a\rangle+c_a}+\frac{n_a|\xi_a|^2_{g_\omega}}{(\langle\mu_\omega,p_a\rangle+c_a)^2}+\frac{s_a}{\langle\mu_\omega,p_a\rangle+c_a}\right).
\end{aligned}
$$

On the other hand, using a basis (ξ_i) of $\mathfrak{t}$ with a dual basis (ξ^i) of $\mathfrak{t}^*$,

$$
\begin{aligned}
\mathrm{Scal}_p(\omega) &:= p(\mu_\omega)\,\mathrm{Scal}(\omega)+2\sum_{i=1}^{r}p_{,i}(\mu_\omega)\Delta_\omega(\langle\mu_\omega,\xi_i\rangle)-\sum_{i,j=1}^{r}p_{,ij}(\mu_\omega)g_\omega(\xi_i,\xi_j) \\
&= p(\mu_\omega)\,\mathrm{Scal}(\omega)+2\sum_{i=1}^{r}\Delta_\omega(\langle\mu_\omega,\xi_i\rangle)\sum_{a=1}^{k}\frac{n_a\xi^i(p_a)p(\mu_\omega)}{\langle\mu_\omega,p_a\rangle+c_a} \\
&\qquad +\sum_{i,j=1}^{r}g_\omega(\xi_i,\xi_j)\left[\sum_{a=1}^{k}\frac{n_a\xi^i(p_a)\xi^j(p_a)p(\mu_\omega)}{(\langle\mu_\omega,p_a\rangle+c_a)^2}\right. \\
&\qquad\qquad \left.-\sum_{a,b=1}^{k}\frac{n_a n_b\xi^i(p_a)\xi^j(p_a)p(\mu_\omega)}{(\langle\mu_\omega,p_a\rangle+c_a)(\langle\mu_\omega,p_b\rangle+c_b)}\right] \\
&= p(\mu_\omega)\,\mathrm{Scal}(\omega)+2\Delta_\omega(\langle\mu_\omega,p_a\rangle)\sum_{a=1}^{k}\frac{n_a p(\mu_\omega)}{\langle\mu_\omega,p_a\rangle+c_a} \\
&\qquad +\left[\sum_{a=1}^{k}\frac{n_a|p_a|^2_{g_\omega}p(\mu_\omega)}{(\langle\mu_\omega,p_a\rangle+c_a)^2}-\sum_{a,b=1}^{k}\frac{n_a n_b g_\omega(p_a,p_b)p(\mu_\omega)}{(\langle\mu_\omega,p_a\rangle+c_a)(\langle\mu_\omega,p_b\rangle+c_b)}\right].
\end{aligned}
$$

Comparing the above expression with (45),

$$\mathrm{Scal}(\tilde{\omega}) = \frac{1}{p(\mu_\omega)}\,\mathrm{Scal}_p(\omega) + \left(\sum_{a=1}^{k} \frac{s_a}{\langle \mu_\omega, p_a\rangle + c_a}\right). \tag{46}$$

Using that (as functions on $X \times P$) $\mu_{\tilde{\omega}} = \mu_\omega$ and $g_\omega(\xi_i,\xi_j) = g_{\tilde{\omega}}(\xi_i,\xi_j)$ (see the proof of Lemma 5.7), we further compute, from (46),

$$\begin{aligned}
\mathrm{Scal}_v(\tilde{\omega}) &= v(\mu_{\tilde{\omega}})\,\mathrm{Scal}(\tilde{\omega}) + 2\sum_{i=1}^{r} v_{,i}(\mu_{\tilde{\omega}})\Delta^Y_{\tilde{\omega}}(\langle \mu_{\tilde{\omega}}, \xi_i\rangle) - \sum_{i,j=1}^{r} v_{,ij}(\mu_{\tilde{\omega}}) g_{\tilde{\omega}}(\xi_i,\xi_j) \\
&= \frac{v(\mu_\omega)}{p(\mu_\omega)}\,\mathrm{Scal}_p(\omega) + v(\mu_\omega)q(\mu_\omega) + 2\sum_{i=1}^{r} v_{,i}(\mu_\omega)\Delta^Y_{\tilde{\omega}}(\langle \mu_\omega, \xi_i\rangle) \\
&\qquad - \sum_{i,j=1}^{r} v_{,ij}(\mu_\omega) g_{\tilde{\omega}}(\xi_i,\xi_j) \\
&= \frac{v(\mu_\omega)}{p(\mu_\omega)}\,\mathrm{Scal}_p(\omega) + v(\mu_\omega)q(\mu_\omega) + 2\sum_{i=1}^{r} v_{,i}(\mu_\omega)\Delta^X_{\omega,p}(\langle \mu_\omega, \xi_i\rangle) \\
&\qquad - \sum_{i,j=1}^{r} v_{,ij}(\mu_\omega) g_{\omega}(\xi_i,\xi_j),
\end{aligned}$$

where, for passing to the last line, we used the identity $\Delta^Y_{\tilde{\omega}} = \Delta^X_{\omega,p}$ established in Lemma A.3. As

$$\Delta^X_{\omega,p}(\psi) := \frac{1}{p(\mu_\omega)}\delta_\omega(p(\mu_\omega)d\psi) = \Delta^X_\omega(\psi) - \sum_{j=1}^{r} \frac{p_{,j}(\mu_\omega)}{p(\mu_\omega)} g_\omega(d\mu^{\xi_j}_\omega, d\psi), \tag{47}$$

we further get

$$\begin{aligned}
&\frac{v(\mu_\omega)}{p(\mu_\omega)}\,\mathrm{Scal}_p(\omega) + 2\sum_{i=1}^{r} v_{,i}(\mu_\omega)\Delta^X_{\omega,p}(\langle \mu_\omega, \xi_i\rangle) - \sum_{i,j=1}^{r} v_{,ij}(\mu_\omega) g_\omega(\xi_i,\xi_j) \\
&\quad = v(\mu_\omega)\,\mathrm{Scal}(\omega) + 2\sum_{i=1}^{r} \frac{v(\mu_\omega)p_{,i}(\mu_\omega)}{p(\mu_\omega)}\Delta^X_\omega(\langle \mu_\omega, \xi_i\rangle) \\
&\qquad - \sum_{i,j=1}^{r} \frac{v(\mu_\omega)p_{,ij}(\mu_\omega)}{p(\mu_\omega)} g_\omega(\xi_i,\xi_j) + 2\sum_{i=1}^{r} v_{,i}(\mu_\omega)\Delta^X_\omega(\langle \mu_\omega, \xi_i\rangle) \\
&\qquad\qquad - 2\sum_{i,j=1}^{r} \frac{v_{,i}(\mu_\omega)p_{,j}(\mu_\omega)}{p(\mu_\omega)} g^X(\xi_i,\xi_j) - \sum_{i,j=1}^{r} v_{,ij}(\mu_\omega) g_\omega(\xi_i,\xi_j)
\end{aligned}$$

$$= \frac{1}{p(\mu_\omega)}\bigg((pv)(\mu_\omega)\operatorname{Scal}_\omega + 2\sum_{i=1}^{r}(pv)_{,i}(\mu_\omega)\Delta^X_\omega(\langle\mu_\omega,\xi_i\rangle) \\ - \sum_{i,j=1}^{r}(pv)_{,ij}(\mu_\omega)g_\omega(\xi_i,\xi_j)\bigg)$$

$$= \frac{1}{p(\mu_\omega)}\operatorname{Scal}_{pv}(\omega).$$

The expression (35) follows from the above formulas. □

Lemma 5.10 *The restriction of the weighted Mabuchi energy $\boldsymbol{M}^Y_{v,w}$ on Y to the subspace $\mathcal{K}_{\mathbb{T}}(X,\omega_0)\subset\mathcal{K}_{\mathbb{T}}(Y,\tilde{\omega}_0)$ is equal to $C\boldsymbol{M}^X_{pv,\tilde{w}}$, where p, w and $\tilde{w}$ are as given in Lemma 5.9 and $C=\operatorname{Vol}(B,\omega_B)$.*

Proof This is a direct corollary of Lemma 5.9 and Definition 1.1. □

We now specialize to the case when each (B_a,ω_a) is a Hodge Kähler–Einstein manifold with positive scalar curvature $s_a = 2n_a k_a$, where $k_a\in\mathbb{N}$. Equivalently, $2\pi c_1(B_a) = k_a[\omega_a]$ for a positive integer k_a and an integral Kähler class $\frac{1}{2\pi}[\omega_a]$. Notice that k_a must be a positive divisor of the Fano index $\operatorname{Ind}(B_a)$ of B_a, which yields the a priori bound $1\le k_a\le\operatorname{Ind}(B_a)$. We also assume that $(X,\mathbb{T})$ is Fano, with canonically normalized momentum polytope Δ. We then have:

Lemma 5.11 *In the setting above, if the affine-linear functions $\langle p_a,\mu\rangle + k_a$ are positive on Δ, then the bundle-compatible Kähler metric $\tilde{\omega}$ on Y corresponding to the constants $c_a = k_a$ belongs to de Rham class $2\pi c_1(Y)$. Furthermore, $\tilde{\omega}$ is a v–soliton if and only if ω is a pv–soliton.*

Proof By using (38) and rearranging the terms in (40), we have the relation (written on Z)

$$\rho_{\tilde{\omega}} = \rho_\omega + \sum_{a=1}^{k}(\langle p_a,\mu_{\rho_\omega}\rangle + c_a)\omega_a + \langle d_X\mu_{\rho_\omega}\wedge\theta\rangle \\ + \sum_{a=1}^{k}(\rho_a - c_a\omega_a) - \tfrac{1}{2}d_Y d_Y^c\log p(\mu_\omega), \tag{48}$$

where $\rho_{\tilde{\omega}}$, ρ_ω and ρ_a denote the Ricci forms of $(Y,\tilde{\omega})$, (X,ω) and (B_a,ω_a), respectively, pulled back to Z, and $\mu_{\rho_\omega} := d_X^c\kappa$ is the "momentum map" with respect to the Ricci form ρ_ω. As in (42), we have $\mu_{\rho_\omega} = \frac{1}{2}\Delta_\omega\mu_\omega$. Suppose $\rho_\omega - \omega = \frac{1}{2}d_X d_X^c h$ for some $\mathbb{T}$–invariant smooth function on X; by using that the momentum polytope Δ is canonically normalized, we have (see (11)) $\mu_{\rho_\omega} - \mu_\omega = d^c h$. A closer look at the proof of Lemma 5.5 and the relation (48) (with $c_a = s_a/(2n_a) = k_a$) show that

$$\rho_{\tilde{\omega}} - \tilde{\omega} = \tfrac{1}{2} d_Y d_Y^c \tilde{h} \quad \text{with } \tilde{h} := h - \log p(\mu_\omega).$$

The claim follows from the above. □

Remark 5.12 Lemma 5.11 provides a useful way to *construct* semisimple $(X, \mathbb{T})$–principal Fano fibrations. Indeed, for given positive Hodge Kähler–Einstein manifolds (B_a, ω_a) as above with corresponding integer constants k_a, and a given Fano manifold $(X, \mathbb{T})$ with associated canonical polytope Δ, one can try to find the possible principal $\mathbb{T}$–bundles P over $B = \prod_{a=1}^k B_a$ for which the corresponding semisimple $(X, \mathbb{T})$–principal fibration is Fano. Such principal $\mathbb{T}$–bundles P are in correspondence with the choice of lattice elements $p_a \in \Lambda \subset \mathfrak{t}$ and Lemma 5.11 tells us that for a set of elements p_a, to determine a Fano semisimple $(X, \mathbb{T})$–principal fibration Y it is sufficient to check that, for all a,

$$\langle p_a, \mu \rangle + k_a > 0 \quad \text{on } \Delta.$$

For instance, if we take $B = B_1 = \mathbb{P}^1$ with a Fubini–Study metric ω_1 of scalar curvature 4 (so that $k_1 = 2$ and ω_1 is primitive) and $(X, \mathbb{T}) = (\mathbb{P}^1, \mathbb{S}^1)$ with canonical polytope $\Delta = [-1, 1]$, then the possible Fano $(\mathbb{P}^1, \mathbb{S}^1)$–principal fibrations will correspond to $p_1 \in \mathbb{Z}$ such that $p_1\mu + 2 > 0$ on $[-1, 1]$, ie $p_1 = \pm 1, 0$ are the only possible values. This gives rise to the Fano surfaces $\mathbb{P}(\mathcal{O} \oplus \mathcal{O}(-1)) \cong \mathbb{P}(\mathcal{O} \oplus \mathcal{O}(1))$ and $\mathbb{P}^1 \times \mathbb{P}^1$. In general, the isomorphism class of the principal $\mathbb{T}$–bundle P over B, and hence also the semisimple $(X, \mathbb{T})$–principal Fano fibration constructed as above, is encoded by the Hodge classes $\frac{1}{2\pi}[\omega_a] \otimes p_a = 1/k_a c_1(B_a) \otimes p_a \in H^2(B, \mathbb{Z})^r$. The a priori bounds $1 \leq k_a \leq \mathrm{Ind}(B_a)$ for k_a show that, for given base $B = \prod_{a=1}^k B_a$ and fibre $(X, \mathbb{T})$, there are only a finite number of semisimple $(X, \mathbb{T})$–principal Fano fibrations constructed this way.

Remark 5.13 The relationship between the Ricci potentials $\tilde{h}$ and h established in the proof of Lemma 5.11 and (29) yield, via Remark 2.3, that, if the momentum map μ_ω of $(X, \omega, \mathbb{T}_X)$ is canonically normalized, then the momentum map $\mu_{\tilde{\omega}} = \mu_\omega$ of the corresponding bundle-compatible Kähler metric $\tilde{\omega}$ on $(Y, \mathbb{T}_Y)$ is also canonically normalized.

We finish this section with a straightforward extension of [6, Lemma 5]:

Lemma 5.14 *Suppose Y is a semisimple principal $(X, \mathbb{T})$–fibration over B, such that $\mathbb{T}$ is a maximal torus in the reduced group of automorphisms $\mathrm{Aut}_r(X)$. Let $\tilde{\omega}$ be a bundle-compatible Kähler metric on Y corresponding to a $\mathbb{T}$–invariant Kähler metric ω on X, and $\mathbb{K}_B \subset \mathrm{Aut}_r(B)$ be a maximal compact torus in the reduced group of automorphisms of B which (without loss of generality by the Lichnerowicz–Matsushima theorem) belongs to the isometry group of ω_B. Then $\tilde{\omega}$ is invariant under the action of a maximal torus $\mathbb{K}_Y \subset \mathrm{Aut}_r(Y)$, and we have an exact sequence*

$$\{0\} \to \mathrm{Lie}(\mathbb{T}_Y) \to \mathrm{Lie}(\mathbb{K}_Y) \to \mathrm{Lie}(\mathbb{K}_B) \to \{0\}.$$

Furthermore, for any positive weight functions v and w_0 defined on $\Delta \subset \mathfrak{t}^*$, there exists a unique affine-linear function $\ell^{\mathrm{ext}}_{v,w_0}$ on $\mathfrak{t}^*$ such that, when pulled back to the dual Lie algebra $\mathfrak{k}_Y^*$ of $\mathbb{K}_Y$, $(v, w_0\ell^{\mathrm{ext}}_{v,w_0})$ satisfies (3) with respect to $\tilde{\omega}$ on Y for any affine-linear function ℓ on $\mathfrak{k}_Y^*$.

Proof This proof is not materially different than the proof of [6, Lemma 5] (which is made in the case when $(X, \mathbb{T})$ is toric and $v = w_0 = 1$). We only give a sketch. A Killing potential f for a Killing vector field $K \in \mathfrak{k}_B := \mathrm{Lie}(\mathbb{K}_B)$ is of the form $f = \sum_{a=1}^k f_a$, where f_a is a Killing potential of (B_a, ω_a). Letting $\tilde{K}$ be the horizontal lift of K to P (using the $\mathfrak{t}_P$–valued connection 1–form θ), one can check that the vector field on P

$$\hat{K} = \tilde{K} + \sum_{a=1}^k f_a \xi^P_{p_a}$$

is a CR vector field on $(P, \mathcal{D}, J_B)$, hence also on $(Z, \mathscr{H}, J_B \oplus J_X)$. Furthermore, a direct verification in (24) reveals that

$$\iota_{\hat{K}} \tilde{\omega} = -d\Big(\sum_{a=1}^k (\langle p_a, \mu_\omega\rangle + c_a) f_a\Big), \tag{49}$$

so $\hat{K}$ also preserves $\tilde{\omega}$. We thus obtain a lift $\hat{\mathfrak{k}}_B$ of the Lie algebra $\mathfrak{k}_B = \mathrm{Lie}(T_B)$ to Z, which clearly commutes with the action $\mathbb{T}_Z$ and preserves both the CR structure of $(Z, \mathscr{H})$ and the 2–form $\tilde{\omega}$. The Lie algebra $\mathfrak{k}_Y$ of $\mathbb{K}_Y$ is then induced by $\mathfrak{t}_X \oplus \hat{\mathfrak{k}}_B \subset TZ$, which descend to an abelian Lie algebra of Killing fields on Y. The maximality of $\mathbb{K}_Y \subset \mathrm{Aut}_r(Y)$ and the exactness of the sequence follow from the maximality of each $\mathbb{K}_B \subset \mathrm{Aut}_r(B)$ and $\mathbb{T} \subset \mathrm{Aut}_r(X)$, and the fact that (recall that Y is a locally trivial X–fibre bundle and therefore the fibres have trivial normal bundle) any holomorphic vector field on Y projects under π_B to a holomorphic vector field on B. For the final claim in Lemma 5.14, notice that by (49) the Killing potentials of all lifted Killing vector fields $\widehat{K}$ from B are of the form $\sum_{a=1}^{k}(\langle p_a,\mu_\omega\rangle + c_a)f_a$. Thus, by Lemma 5.9 and using (27), the integral condition (3) on $(Y,\tilde{\omega})$ will be zero for any such Killing potential as soon as we normalize $\int_{B_a} f_a\omega_a^{n_a}=0$ and assume $\ell^{\mathrm{ext}}_{v,w_0}\in \mathrm{Aff}(\mathfrak{t}^*_X)$. On the other hand, examining (3) on $(Y,\tilde{\omega})$ for the Killing potentials $\ell(\mu_{\tilde{\omega}})$ for $\ell\in\mathrm{Aff}(\mathfrak{t}^*)$ reduces (again by Lemma 5.9 and (29)) to an integral relation on (X,ω) which defines a unique element $\ell^{\mathrm{ext}}_{v,w_0}\in\mathrm{Aff}(\mathfrak{t}^*)$. □

6 Weighted functionals and distances and their extensions

Let ω_0 be a $\mathbb{T}$–invariant Kähler metric in the Kähler class α, denote by $\mathrm{PSH}_{\mathbb{T}}(X,\omega_0)$ the space of $\mathbb{T}$–invariant ω_0–plurisubharmonic functions in $L^1(X,\omega_0)$, and define the class of potentials of full volume by

$$\mathcal{E}_{\mathbb{T}}(X,\omega_0) := \left\{\varphi\in\mathrm{PSH}_{\mathbb{T}}(X,\omega_0)\;\middle|\;\int_X \mathrm{MA}(\varphi)=\int_X\omega_0^{[n]}\right\}.$$

According to [27], the d_1–completion of $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ can be identified with the subspace of potentials of finite energy:

$$\mathcal{E}^1_{\mathbb{T}}(X,\omega_0) = \left\{\varphi\in\mathcal{E}_{\mathbb{T}}(X,\omega_0)\;\middle|\;\int_X|\varphi|\,\mathrm{MA}(\varphi)<\infty\right\}.$$

Our main result in this section will be the existence of an lsc extension of the weighted Mabuchi functional (defined in Definition 1.1 on the space $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$) to a functional on $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$. Our starting point is that the weighted Mabuchi energy $\boldsymbol{M}_{v,w}$ admits a weighted Chen–Tian decomposition [55, Theorem 5] into energy and entropy parts

$$\text{(50)}\quad \boldsymbol{M}_{v,w}(\varphi) = \int_X \log\left(\frac{v(\mu_\varphi)\omega_\varphi^m}{\omega_0^m}\right)v(\mu_\varphi)\omega_\varphi^{[m]} - 2\boldsymbol{I}_v^{\rho_{\omega_0}}(\varphi)+\boldsymbol{I}_w(\varphi) - \int_X\log(v(\mu_0))v(\mu_0)\omega_0^{[m]},$$

where ρ_{ω_0} is the Ricci form of ω_0 and the functionals $\boldsymbol{I}_w$ and $\boldsymbol{I}_v^{\rho_{\omega_0}}$ are introduced in Definition 6.2. We want to show:

Theorem 6.1 *For smooth weight functions $v(\mu)$ and $w(\mu)$ such that $v(\mu)>0$ on Δ, the weighted Mabuchi energy $\boldsymbol{M}_{v,w}\colon\mathcal{K}_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}$ extends using* (50) *to the largest d_1–lsc functional $\boldsymbol{M}_{v,w}\colon\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}\cup\{\infty\}$ which is convex along the finite-energy geodesics of $\mathcal{E}_{\mathbb{T}}(X,\omega_0)$. Additionally, the extended weighted Mabuchi energy $\boldsymbol{M}_{v,w}$ is linear in v and w, uniformly continuous in w in the $C^0(\Delta)$ topology, and continuous with respect to v in the $C^1(\Delta)$ topology.*

The above result is well known for the unweighted case, by [13], and we will follow a similar path to get an extension in the weighted case. The proof of Theorem 6.1 will be given at the end of the section, and we detail below the definition and extension of each component of (50).

6.1 The weighted Aubin–Mabuchi functionals

Definition 6.2 [55] For a smooth weight function $v(\mu)$ on Δ we let $\boldsymbol{I}_v$ denote the functional on $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ defined by

$$(d_\varphi \boldsymbol{I}_v)(\dot\varphi) = \int_X \dot\varphi v(\mu_\varphi)\omega_\varphi^{[n]}, \quad \boldsymbol{I}_v(0)=0,$$

and let $\boldsymbol{J}_v := \int_X \varphi v(\mu_0)\omega_0^{[m]} - \boldsymbol{I}_v(\varphi)$. Furthermore, for a fixed $\mathbb{T}$–invariant closed $(1,1)$–form ρ on X with momentum $\mu_\rho\colon X \to \mathfrak{t}^*$, we define the *$\rho$–twisted Aubin–Mabuchi functional* $\boldsymbol{I}_v^\rho\colon \mathcal{K}_{\mathbb{T}}(X,\omega_0) \to \mathbb{R}$ by

$$(d_\varphi \boldsymbol{I}_v^\rho)(\dot\varphi) := \int_X \dot\varphi\big(v(\mu_\varphi)\rho\wedge\omega_\varphi^{[m-1]} + \langle (dv)(\mu_\varphi),\mu_\rho\rangle\omega_\varphi^{[m]}\big), \quad \boldsymbol{I}_v^\rho(0)=0.$$

For $v\equiv 1$, we let $\boldsymbol{I}_1=\boldsymbol{I}$, $\boldsymbol{J}_1=\boldsymbol{J}$ and $\boldsymbol{I}_v^\rho=\boldsymbol{I}^\rho$, and notice that $\boldsymbol{I}$ and $\boldsymbol{J}$ are the functionals introduced in Definition 3.1

Remark 6.3 It follows from the above definition and the results in [55] that for any weight $v(x)$ and a constant c, $\boldsymbol{J}_v(\varphi+c)=\boldsymbol{J}_v(\varphi)$, allowing one to see $\boldsymbol{J}_v$ as a functional on the space of $\mathbb{T}$–invariant Kähler metrics in the Kähler class $\alpha=[\omega_0]$, and motivating the notation $\boldsymbol{J}_v(\omega_\varphi)$. Notice also that $\boldsymbol{I}_v$, $\boldsymbol{J}_v$ and $\boldsymbol{I}_v^\rho$ are linear in v. In the case when $v>0$, $\boldsymbol{J}_v$ is nonnegative (see Lemma 6.4), whereas $\boldsymbol{I}_v$ is monotone in the sense that, for any $\varphi_0,\varphi_1 \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ with $\varphi_1(x)\ge\varphi_0(x)$,

$$\boldsymbol{I}_v(\varphi_1)-\boldsymbol{I}_v(\varphi_0) \ge \inf_\Delta(v)\int_X(\varphi_1-\varphi_0)\omega_{\varphi_0}^{[m]}.$$

The above inequality follows by Definition 6.2, integrating the derivative of $\boldsymbol{I}_v$ along the path $t\varphi_1+(1-t)\varphi_0 \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$, and integrating by parts.

The following is established in [45, (2.37)]:

Lemma 6.4 *Let $v>0$. There exists a uniform constant $C=C(X,\omega_0,v)>0$ such that*

$$\frac{1}{C}\boldsymbol{J}(\varphi) \le \boldsymbol{J}_v(\varphi) \le C\boldsymbol{J}(\varphi).$$

Proof Let $\varphi_t := \varphi_0 + t\varphi$ with $\varphi := \varphi_1 - \varphi_0$ and $\omega_{\varphi_t} = \omega_{\varphi_0} + t dd^c\varphi$ for $t \in [0,1]$. We compute

$$\begin{aligned}
\boldsymbol{J}_v(\varphi) = \boldsymbol{J}_v(\varphi_1) - \boldsymbol{J}_v(\varphi_0) &= \int_0^1 \int_X \varphi(v(\mu_\omega)\omega^{[m]} - v(\mu_{\varphi_s})\omega_{\varphi_s}^{[m]})\,ds \\
&= -\int_0^1 \int_X \varphi\Big(\int_0^s \frac{d}{dt}[v(\mu_{\varphi_t})\omega_{\varphi_t}^{[m]}]\,dt\Big)\,ds \\
&= -\int_0^1 \int_X \varphi\Big(\int_0^s \big(g_{\varphi_t}(d[\log\circ v(\mu_{\varphi_t})], d\varphi) - \Delta_{\varphi_t}(\varphi)\big) v(\mu_{\varphi_t})\omega_{\varphi_t}^{[m]}\,dt\Big)\,ds \\
&= -\int_0^1 \int_0^s \Big(\int_X \varphi d[v(\mu_{\varphi_t})] \wedge d^c\varphi \wedge \omega_{\varphi_t}^{[m-1]} + \varphi dd^c\varphi \wedge v(\mu_{\varphi_t})\omega_{\varphi_t}^{[m-1]}\Big)\,dt\,ds \\
&= \int_0^1 \int_0^s \Big(\int_X v(\mu_{\varphi_t}) d\varphi \wedge d^c\varphi \wedge \omega_{\varphi_t}^{[m-1]}\Big)\,dt\,ds \\
&= \int_0^1 \int_0^s \Big(\int_X v(\mu_{\varphi_t}) d\varphi \wedge d^c\varphi \wedge (t\omega_\varphi + (1-t)\omega)^{[m-1]}\Big)\,dt\,ds \\
&= \sum_{j=0}^{m-1} \int_0^1 \int_0^s \Big(\int_X t^j(1-t)^{m-j-1} v(\mu_{\varphi_t}) d\varphi \wedge d^c\varphi \wedge \omega^{[j]} \wedge \omega_\varphi^{[m-j-1]}\Big)\,dt\,ds,
\end{aligned}$$

where, in the fourth equality, we have used that

$$\frac{d}{dt}[v(\mu_{\varphi_t})] = \sum_{i=1}^r v_{,i}(\mu_{\varphi_t})(d^c\varphi)(\xi_i) = g_{\varphi_t}(d[v(\mu_{\varphi_t})], d\varphi) \tag{51}$$

for any basis $(\xi_i)_{i=1,\dots,r}$ of $\mathfrak{t}$. It follows that

$$\frac{1}{C}\boldsymbol{J}(\varphi) \le \boldsymbol{J}_v(\varphi) \le C\boldsymbol{J}(\varphi),$$

where $C = C(X,\alpha,v)$ is a constant such that $1/C \leqslant v \leqslant C$ on Δ_α. □

Lemma 6.5 *Suppose v and w are smooth functions on Δ. Then*

$$\begin{aligned}
|\boldsymbol{J}_v(\varphi) - \boldsymbol{J}_w(\varphi)| &\le \|v-w\|_{C^0(\Delta)} \boldsymbol{J}_1(\varphi), \\
|\boldsymbol{I}_v(\varphi) - \boldsymbol{I}_w(\varphi)| &\le \|v-w\|_{C^0(\Delta)} (\|\varphi\|_{L^1(X,\omega_0)} + \boldsymbol{J}_1(\varphi)).
\end{aligned}$$

In particular, for a fixed $\varphi \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$, $\boldsymbol{I}_v(\varphi)$ and $\boldsymbol{J}_v(\varphi)$ are uniformly continuous in v.

Proof The first relation follows from Lemma 6.4 whereas the second inequality follows from the first and Definition 6.2. □

Lemma 6.6 *The restrictions of $\boldsymbol{I}_1^Y$ and $\boldsymbol{J}_1^Y$ to the subspace $\mathcal{K}_{\mathbb{T}}(X,\omega_0) \subset \mathcal{K}_{\mathbb{T}}(Y,\tilde{\omega}_0)$ are equal to $C\boldsymbol{I}_p^X$ and $C\boldsymbol{J}_p^X$, respectively, where $p(\mu)$ is the weight function defined* *in Lemma 5.9 and* $C = \mathrm{Vol}(B, \omega_B)$. *Furthermore, if* $\tilde{\rho}$ *is a Kähler form on* Y *induced by a Kähler form* ρ *on* X *using* (24), *then the restriction of* $(\boldsymbol{I}_1^{\tilde{\rho}})^Y$ *to the subspace* $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$ *equals* $C(\boldsymbol{I}_p^{\rho})^X$.

Proof The first part follows from the definition of $\boldsymbol{I}_1^Y$, using that

$$\tilde{\omega}_\varphi^{[n+m]} \wedge \theta^{\wedge r} = p(\mu_\varphi)\omega_\varphi^{[m]} \wedge \omega_{\boldsymbol{B}}^{[n]} \wedge \theta^{\wedge r}$$

on Z.

Similarly, if $\tilde{\rho}$ is a $(1,1)$–form Y whose pullback to Z is

$$\tilde{\rho} := \rho + \sum_{a=1}^{k} (\langle p_a, \mu_\rho \rangle + c_a))\pi_{\boldsymbol{B}}^* \omega_a + \langle d\mu_\rho \wedge \theta \rangle, \tag{52}$$

we compute

$$\begin{aligned}(d_\varphi \boldsymbol{I}_p^\rho)^X(\dot{\varphi}) &= \int_X \dot{\varphi}[p(\mu_\varphi)\rho \wedge \omega_\varphi^{[m-1]} + \langle (dp)(\mu_\varphi), \mu_\rho \rangle \omega_\varphi^{[m]}] \\ &= \frac{1}{\mathrm{Vol}(B, \omega_{\boldsymbol{B}})} \int_Y \dot{\varphi}\tilde{\rho} \wedge \tilde{\omega}_\varphi^{[n+m-1]} = \frac{1}{\mathrm{Vol}(B, \omega_{\boldsymbol{B}})} (d_\varphi \boldsymbol{I}^{\tilde{\rho}})^Y(\dot{\varphi}).\end{aligned}$$

The claim follows, as $(\boldsymbol{I}_p^\rho)^X(0) = 0 = (\boldsymbol{I}^{\tilde{\rho}})^Y(0)$. □

6.2 The weighted d_1–distance

Definition 6.7 Let $v > 0$ be a positive function on Δ. For $\varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ we let

$$d_{1,v}(\varphi_0, \varphi_1) := \inf_{\psi(t)} \{L_{1,v}(\psi(t)) \mid \psi(t,x) \in C_{\mathbb{T}}^\infty([0,1] \times X) \text{ and } \psi(t) \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)\},$$

where

$$L_{1,v}(\psi(t)) := \int_0^1 \left(\int_X |\dot{\psi}(t)| v(\mu_{\psi(t)}) \omega_{\psi(t)}^{[m]} \right) dt.$$

For $v \equiv 1$, we have $d_{1,1} = d_1$, where d_1 is the distance introduced in Section 3.

Lemma 6.8 *For any weight* $v > 0$, *there exists uniform constant* $C = C(X, \omega_0, v) > 0$ *such that*

$$\frac{1}{C} d_1(\varphi_0, \varphi_1) \le d_{1,v}(\varphi_0, \varphi_1) \le C d_1(\varphi_0, \varphi_1) \quad \text{for all } \varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0), \tag{53}$$

where $d_1 := d_{1,1}$ *is the distance introduced in* [27]. *In particular,* $d_{1,v}$ *is a distance on* $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$ *which is quasiisometric with* d_1.

Proof The relation (53) follows from the fact that $v(\mu)$ is positive and uniformly bounded on Δ. This yields that $d_{1,v}$ is a distance, as d_1 is a distance according to [27]. □

Lemma 6.9 *For any smooth weight* $v > 0$,

$$|\boldsymbol{I}_v(\varphi_0) - \boldsymbol{I}_v(\varphi_1)| \leq d_{1,v}(\varphi_0, \varphi_1) \leq C d_1(\varphi_0, \varphi_1) \quad \textit{for all } \varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0).$$

Proof For any smooth curve φ_t between φ_1 and φ_2, using Definition 6.2,

$$|\boldsymbol{I}_v(\varphi_0) - \boldsymbol{I}_v(\varphi_1)| = \left| \int_0^1 (d_{\varphi_t} \boldsymbol{I}_v)(\dot{\varphi}_t)\, dt \right| \leq L_{1,v}(\varphi_t).$$

The claim follows from the above and Lemma 6.8. □

6.3 Extensions to $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$

Lemma 6.10 *For any smooth weight* v, *the functionals* $\boldsymbol{I}_v$ *and* $\boldsymbol{J}_v$ *continuously extend to the space* $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$. *Furthermore, for any* $\psi \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$, *the extended functionals are linear and uniformly continuous in* v, *in the topology* $C^0(\Delta)$.

Proof $\boldsymbol{I}_v$ is d_1–Lipschitz by Lemma 6.9; for $\boldsymbol{J}_v$, we get from Definition 6.2 that

$$|\boldsymbol{J}_v(\varphi_0) - \boldsymbol{J}_v(\varphi_1)| \leq \int_X |\varphi_0 - \varphi_1| \omega_0^{[m]} + |\boldsymbol{I}_v(\varphi_0) - \boldsymbol{I}_v(\varphi_1)|.$$

Combining the above inequality with Lemma 6.9 and [27, Corollary 5.7], there exists a uniform positive constant $C = C(X, \omega_0, v)$ and, for any fixed positive real number $R > 0$, an increasing continuous function $F_R \colon \mathbb{R}_+ \to \mathbb{R}_+$, $F(0) = 0$, defined in terms of (X, ω_0, R), such that, for any $\varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ with $d_1(0, \varphi_i) \leq R$,

$$|\boldsymbol{J}_v(\varphi_0) - \boldsymbol{J}_v(\varphi_1)| \leq C d_1(\varphi_0, \varphi_1) + F_R(d_1(\varphi_0, \varphi_1)),$$

showing that $\boldsymbol{J}_v$ is locally uniformly continuous on $(\mathcal{K}_{\mathbb{T}}(X, \omega_0), d_1)$ and thus extends continuously to $(\mathcal{E}^1_{\mathbb{T}}(X, \omega_0), d_1)$.

The v–linearity of $\boldsymbol{I}_v$ and $\boldsymbol{J}_v$ is clear by continuity; see Remark 6.3. The continuity with respect to v follows from the continuous extensions of the inequalities in Lemma 6.5, noting that we have already shown that $\boldsymbol{J}_v$, $\boldsymbol{J}_w$, $\boldsymbol{J}$, $\boldsymbol{I}_v$ and $\boldsymbol{I}_w$ all extend continuously, whereas $\|\cdot\|_{L^1(X,\omega_0)}$ extends continuously by [27, Theorem 5.8]. □

Corollary 6.11 *The metric completion of* $(\mathcal{K}_{\mathbb{T}}(X, \omega_0) \cap \boldsymbol{I}_v^{-1}(0), d_1)$ *is the complete geodesic metric space* $(\mathcal{E}^1_{\mathbb{T}}(X, \omega_0) \cap \boldsymbol{I}_v^{-1}(0), d_1)$.

Proof Similarly to [29, Lemma 5.2], one can show that $\boldsymbol{I}_v$ is linear along finite-energy geodesics. As $\boldsymbol{I}_v \colon \mathcal{E}^1_{\mathbb{T}}(X, \omega_0) \to \mathbb{R}$ is d_1–continuous, it follows that $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0) \cap \boldsymbol{I}_v^{-1}(0)$ is a d_1–closed subspace. □

Lemma 6.12 *Let v be a smooth weight function and ρ a $\mathbb{T}$–invariant closed $(1,1)$–form. The functional $\boldsymbol{I}_v^{\rho}\colon \mathcal{K}_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}$ extends to a d_1–continuous functional on $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ which is bounded on d_1–bounded subsets of $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$. Furthermore, the extended functional is linear and uniformly continuous in v, in the $C^1(\Delta)$ topology.*

Proof Following the proof of [14, Proposition 4.4], we show that $\boldsymbol{I}_v^{\rho}$ is locally uniformly d_1–continuous and bounded on d_1–bounded subsets of $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$. Letting $\varphi_0,\varphi_1\in\mathcal{K}_{\mathbb{T}}(X,\omega_0)$, we put $\varphi_s := s\varphi_1+(1-s)\varphi_0$ for $s\in[0,1]$ and compute

$$
\begin{aligned}
\text{(54)}\quad \boldsymbol{I}_v^{\rho}(\varphi_1)-\boldsymbol{I}_v^{\rho}(\varphi_0) &= \int_0^1 \frac{d}{ds}\boldsymbol{I}_v^{\rho}(\varphi_s)\,ds \\
&= \int_0^1\int_X(\varphi_1-\varphi_0)\big(v(\mu_{\varphi_s})\rho\wedge\omega_{\varphi_s}^{[m-1]}+\langle(dv)(\mu_{\varphi_s}),\mu_\rho\rangle\omega_{\varphi_s}^{[m]}\big)\,ds \\
&= \int_X(\varphi_1-\varphi_0)\sum_{j=0}^{m-1} v_{j,m-1}(\mu_{\varphi_0},\mu_{\varphi_1})\rho\wedge\omega_{\varphi_1}^{[j]}\wedge\omega_{\varphi_0}^{[m-j-1]} \\
&\quad + \int_X(\varphi_1-\varphi_0)\sum_{j=0}^{m}\langle(dv)_{j,m}(\mu_{\varphi_0},\mu_{\varphi_1}),\mu_\rho\rangle\omega_{\varphi_1}^{[j]}\wedge\omega_{\varphi_0}^{[m-j]},
\end{aligned}
$$

where $v_{j,k}(\mu_0,\mu_1)$ and $(dv)_{j,k}(\mu_0,\mu_1)$ are defined on $\Delta\times\Delta$ by

$$
\begin{aligned}
v_{j,k}(\mu_0,\mu_1) &:= \int_0^1 s^j(1-s)^{k-j}v(s\mu_1+(1-s)\mu_0), \\
(dv)_{j,k}(\mu_0,\mu_1) &= \int_0^1 s^j(1-s)^{k-j}(dv)(s\mu_1+(1-s)\mu_0).
\end{aligned}
$$

Using the computation (54),

$$
\begin{aligned}
\text{(55)}\quad |\boldsymbol{I}_v^{\rho}(\varphi_1)-\boldsymbol{I}_v^{\rho}(\varphi_0)| &\leqslant C\int_X|\varphi_1-\varphi_0|\sum_{j=0}^{m-1}\omega_0\wedge\omega_{\varphi_1}^{[j]}\wedge\omega_{\varphi_0}^{[m-j-1]} \\
&\quad + C\int_X|\varphi_1-\varphi_0|\sum_{j=0}^{m}\omega_{\varphi_1}^{[j]}\wedge\omega_{\varphi_0}^{[m-j]} \\
&\leqslant C\int_X|\varphi_1-\varphi_0|\omega_{(\varphi_0+\varphi_1)/4}^{[m]},
\end{aligned}
$$

where in the first inequality we use that the functions $\langle(dv)_{j,k}(\mu_{\varphi_0},\mu_{\varphi_1}),\mu_\rho\rangle$ and $v_{j,k}(\mu_{\varphi_0},\mu_{\varphi_1})$ are bounded on $\Delta\times\Delta$ and $-C\omega_0<\rho<C\omega_0$ for some constant $C>1$, and in the second inequality we use the observation $\omega_{(\varphi_0+\varphi_1)/4}=\frac12\omega_0+\frac14\omega_{\varphi_0}+\frac14\omega_{\varphi_1}$. Using the estimate (55) we can show, similarly to [14, Proposition 4.4], that for any $R>0$ there is an increasing continuous function $F_R\colon\mathbb{R}\to\mathbb{R}$ with $F_R(0)=0$ such that

$$|\boldsymbol{I}_v^{\rho}(\varphi_1)-\boldsymbol{I}_v^{\rho}(\varphi_0)|\leqslant F_R(d_1(\varphi_0,\varphi_1))$$

for any $\varphi_0, \varphi_1 \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \cap \{\varphi \mid d_1(0, \varphi) < R\}$. It follows that $\boldsymbol{I}_v^{\rho}$ extends to a d_1–continuous functional on $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ which is bounded on d_1–bounded subsets of $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$.

For the last statement, let v and w be two (smooth) positive weight functions and $\varphi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$. Taking $\varphi_1 = \varphi$ and $\varphi_0 = 0$ in the computation (54),

$$\boldsymbol{I}_v^{\rho}(\varphi) = \int_X \varphi \sum_{j=0}^{m-1} v_{j,m-1}(\mu_0, \mu_\varphi)\rho \wedge \omega_\varphi^{[j]} \wedge \omega_0^{[m-j-1]} + \int_X \varphi \sum_{j=0}^{m} \langle (dv)_{j,m}(\mu_0, \mu_\varphi), \mu_\rho \rangle \omega_\varphi^{[j]} \wedge \omega_0^{[m-j]}.$$

Let $C > 1$ such that $-C\omega_0 < \rho < C\omega_0$. Using the above formula,

$$\begin{aligned} |\boldsymbol{I}_v^{\rho}(\varphi) - \boldsymbol{I}_w^{\rho}(\varphi)| &= |\boldsymbol{I}_{v-w}^{\rho}(\varphi)| \\ &\leqslant C \int_X |\varphi| \sum_{j=0}^{m-1} |(v-w)_{j,m-1}(\mu_0, \mu_\varphi)| \omega_\varphi^{[j]} \wedge \omega_0^{[m-j]} \\ &\quad + C \int_X |\varphi| \sum_{j=0}^{m} |\langle (d(v-w))_{j,m}(\mu_0, \mu_\varphi), \mu_\rho \rangle| \omega_\varphi^{[j]} \wedge \omega_0^{[m-j]} \\ &\leqslant C \|v-w\|_{C^1(\Delta)} \int_X \sum_{j=0}^{m} |\varphi| \omega_\varphi^{[j]} \wedge \omega_0^{[m-j]} \\ &\leqslant C \|v-w\|_{C^1(\Delta)} \int_X |\varphi| (2\omega_0 + dd^c\varphi)^{[m]} \\ &\leqslant C \|v-w\|_{C^1(\Delta)} \int_X |\varphi| \omega_\varphi^{[m]}. \end{aligned}$$

Using approximation by decreasing sequences in $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$, the above estimate holds for $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$. □

Following Berman and Nyström [15] and the recent work of Han and Li [45], we now define the extension of weighted Monge–Ampère measures to the space $\mathcal{E}_{\mathbb{T}}(X, \omega_0)$.

Proposition 6.13 *Let $v > 0$ be a smooth weight function. For any $\varphi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$, let*

$$\mathrm{MA}_v(\varphi) := v(\mu_\varphi)\omega_\varphi^{[m]}.$$

Then $\mathrm{MA}_v(\varphi)$ extends to a well-defined Radon measure defined for any $\varphi \in \mathcal{E}_{\mathbb{T}}(X, \omega_0)$ such that, for any decreasing sequence $(\varphi_j)_j$ of elements in $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$ converging to φ (which exists by [17]), *we have $\lim_{j\to\infty} \mathrm{MA}_v(\varphi_j) = \mathrm{MA}_v(\varphi)$.*

Proof The result is established in [15; 45] for $\omega_0 \in \alpha = c_1(L)$ a Kähler Hodge class on a projective variety X. The method of Han and Li [45, Proposition 2.2], which uses the semisimple principal fibration construction and polynomial approximations, extends to the case of an arbitrary Kähler class $\alpha = [\omega_0]$. Below we give details of this construction, for the reader's convenience.

Let $\varphi \in \mathcal{E}_{\mathbb{T}}(X, \omega_0)$. Following the proof of [45, Proposition 2.2], we first define $\mathrm{MA}_p(\varphi)$ for a positive polynomial weight of the form $p(\mu) := \prod_{a=1}^k (\langle p_a, \mu\rangle + c_a)^{n_a}$, and extend the definition linearly on p for finite sums of such polynomials. We can then use the Bernstein approximation theorem of an arbitrary positive v with polynomials of the above form in order to obtain $\mathrm{MA}_v(\varphi)$.

We start with a semisimple principal $(X, \mathbb{T})$–fibration Y (see Section 5) with corresponding polynomial weight $p(\mu) := \prod_{a=1}^k (\langle p_a, \mu\rangle + c_a)^{n_a}$; see (28). As the choice of the base $B = B_1 \times \cdots \times B_k$ does not matter, we can simply take (as in [45]) B to be the product of projective spaces $(B_a, \omega_a) = (\mathbb{P}^{n_a}, \omega_a)$ endowed with Fubini–Study metrics of scalar curvatures $2n_a(n_a + 1)$, and P to be the principal $\mathrm{U}(1)^r$–bundle over B, obtained from the tensor products P_i of (the pullbacks to B of) the natural principal $\mathrm{U}(1)$–bundles of degrees p_{ai} over $\mathbb{P}^{n_a}$; see Remark 5.1.

Using [17, Theorem 1], there is a decreasing sequence

$$\varphi_j \in \mathrm{PSH}_{\mathbb{T}}(X, \omega_0) \cap C^\infty(X) = \mathcal{K}_{\mathbb{T}}(X, \omega_0)$$

converging towards φ. By Lemma 5.5 we have $\varphi_j \in \mathcal{K}_{\mathbb{T}}(Y, \tilde{\omega}_0)$ and, by (29), for any $\mathbb{T}_X$–invariant continuous function f on X,

$$\int_X f p(\mu_{\varphi_j})(\omega_0 + d_X d_X^c \varphi_j)^{[m]} = \frac{1}{\mathrm{Vol}(B, \omega_B)} \int_Y f(\tilde{\omega}_0 + d_Y d_Y^c \varphi_j)^{[m+n]}.$$

Passing to the limit in both sides of the above equation, we can define $\mathrm{MA}_p^X(\varphi)$ on $\mathbb{T}$–invariant continuous functions f by

$$\int_X f\, \mathrm{MA}_p^X(\varphi) := \lim_{j\to\infty} \frac{1}{\mathrm{Vol}(B, \omega_B)} \int_Y f(\tilde{\omega}_0 + d_Y d_Y^c \varphi_j)^{[m+n]}. \tag{56}$$

Notice that by [43, Theorem 1.9] the limit exists and is well defined on Y (independent of the chosen sequence).

For a continuous function f on X which is not necessarily $\mathbb{T}_X$–invariant, we define

$$\int_X f\, \mathrm{MA}_p^X(\varphi) := \int_X f^{\mathbb{T}}\, \mathrm{MA}_p^X(\varphi),$$

where $f^{\mathbb{T}}$ is the $\mathbb{T}_X$–invariant function given by the average of f over the $\mathbb{T}_X$–action. It follows that $\mathrm{MA}_p^X(\varphi)$ is a well-defined Radon measure by the Riesz representation theorem.

We can extend the above definition by linearity in p on polynomials which are linear combinations with positive coefficients of polynomials of the above special form. Thus, for $\varphi \in \mathrm{PSH}_{\mathbb{T}}(X, \omega_0)$ and for two polynomials p and q on Δ,

$$\left| \int_X f\,\mathrm{MA}_p^X(\varphi) - \int_X f\,\mathrm{MA}_q^X(\varphi) \right| \leqslant \|p-q\|_{C^0(\Delta)} \int_X |f|\,\mathrm{MA}^X(\varphi) \tag{57}$$

for any $f \in C^0(X)$.

For an arbitrary smooth positive function v on Δ we can approximate v in $C^0(\Delta)$ by polynomials p_i as above (eg by using Bernstein's approximation theorem), and thus, for any continuous function f, the limit

$$\lim_{i\to\infty} \lim_{j\to\infty} \int_X f\,\mathrm{MA}_{p_i}^X(\varphi_j)$$

exists independently of the chosen approximation. We then define

$$\int_X f\,\mathrm{MA}_v^X(\varphi) := \lim_{i\to\infty} \lim_{j\to\infty} \int_X f\,\mathrm{MA}_{p_i}^X(\varphi_j).$$

By the Riesz representation theorem, $\mathrm{MA}_v^X(\varphi)$ is a well-defined Radon measure. □

Remark 6.14 For any $\varphi \in \mathcal{E}_{\mathbb{T}}(X, \omega_0)$, the measure $\mathrm{MA}_v(\varphi)$ is absolutely continuous with respect to $\mathrm{MA}(\varphi)$ since v is bounded on Δ. In particular, for any positive weight v,

$$\mathcal{E}^1_{\mathbb{T}}(X, \omega_0) = \left\{ \varphi \in \mathcal{E}_{\mathbb{T}}(X, \omega_0) \;\middle|\; \int_X |\varphi|\,\mathrm{MA}_v(\varphi) < \infty \right\}.$$

Lemma 6.15 *Let v be a positive weight function and $\varphi_j, \varphi \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ such that $d_1(\varphi_j, \varphi) \to 0$. Then $\mathrm{MA}_v(\varphi_j) \to \mathrm{MA}_v(\varphi)$ weakly.*

Proof Let $v(\mu)$ be a polynomial of the form $p(\mu) := \prod_{a=1}^k (\langle p_a, \mu\rangle + c_a)^{n_a}$ for $\varphi_j \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$, and f any continuous $\mathbb{T}$–invariant function on X. We then have, by the construction in Section 5,

$$\int_X f p(\mu_{\varphi_j})(\omega_0 + d_X d_X^c \varphi_j)^{[m]} = \frac{1}{\mathrm{Vol}(B, \omega_B)} \int_Y f(\tilde{\omega}_0 + d_Y d_Y^c \varphi_j)^{[m+n]}.$$

It follows that, for each $\varphi_j \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ (using an approximation with a decreasing sequence of smooth potentials [17]),

$$\int_X f\,\mathrm{MA}_p^X(\varphi_j) = \frac{1}{\mathrm{Vol}(B, \omega_B)} \int_Y f\,\mathrm{MA}^Y(\varphi_j).$$

By [27, Theorem 5], $\mathrm{MA}^Y(\varphi_j) \to \mathrm{MA}^Y(\varphi)$ weakly as $j \to \infty$. It follows that

$$\lim_{j\to\infty} \int_X f\,\mathrm{MA}_p^X(\varphi_j) = \frac{1}{\mathrm{Vol}(B,\omega_B)} \int_Y f\,\mathrm{MA}^Y(\varphi) = \int_X f\,\mathrm{MA}_p^X(\varphi).$$

Using (56), we conclude that $\mathrm{MA}_p^X(\varphi_j) \to \mathrm{MA}_p^X(\varphi)$ weakly as $j \to \infty$.

For an arbitrary weight function $v \in C^0(\Delta)$, we take a sequence of polynomials p_i of the above form converging to v in $C^0(\Delta)$. For any continuous function f on X, using (57),

$$\begin{aligned}
&\left|\int_X f\,\mathrm{MA}_v(\varphi_j) - \int_X f\,\mathrm{MA}_v(\varphi)\right| \\
&\quad\leqslant \left|\int_X f\,\mathrm{MA}_v(\varphi_j) - \int_X f\,\mathrm{MA}_{p_i}(\varphi_j)\right| + \left|\int_X f\,\mathrm{MA}_{p_i}(\varphi_j) - \int_X f\,\mathrm{MA}_{p_i}(\varphi)\right| \\
&\qquad\qquad + \left|\int_X f\,\mathrm{MA}_{p_i}(\varphi) - \int_X f\,\mathrm{MA}_v(\varphi)\right| \\
&\quad\leqslant \left|\int_X f\,\mathrm{MA}_{p_i}(\varphi_j) - \int_X f\,\mathrm{MA}_{p_i}(\varphi)\right| \\
&\qquad\qquad + \|p_i - v\|_{C^0(\Delta)}\left(\int_X |f|\,\mathrm{MA}(\varphi_j) + \int_X |f|\,\mathrm{MA}(\varphi)\right).
\end{aligned}$$

Letting $j \to \infty$,

$$\lim_{j\to\infty}\left|\int_X f\,\mathrm{MA}_v(\varphi_j) - \int_X f\,\mathrm{MA}_v(\varphi)\right| \leqslant 2\|p_i - v\|_{C^0(\Delta)} \int_X |f|\,\mathrm{MA}(\varphi),$$

using the existence of the weak limits $\mathrm{MA}_{p_i}(\varphi_j) \to \mathrm{MA}_{p_i}(\varphi)$ and $\mathrm{MA}(\varphi_j) \to \mathrm{MA}(\varphi)$ as $j \to \infty$ (by [27, Theorem 5]). Taking the limit $i \to \infty$ in the above inequality,

$$\lim_{j\to\infty}\left|\int_X f\,\mathrm{MA}_v(\varphi_j) - \int_X f\,\mathrm{MA}_v(\varphi)\right| = 0.$$

It follows that $\mathrm{MA}_v(\varphi_j) \to \mathrm{MA}_v(\varphi)$ weakly as $j \to \infty$. □

For a finite measure χ on X we define the entropy of χ with respect to $\omega^{[m]}$ by

$$\mathrm{Ent}(\omega^{[m]}, \chi) := \int_X \log\left(\frac{\chi}{\omega^{[m]}}\right)\chi.$$

In the following lemma we show that the elements of $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ can be approximated in the d_1 distance by smooth potentials with converging entropy of the corresponding weighted Monge–Ampère measures. This is the weighted analogue of [14, Lemma 3.1].

Lemma 6.16 *If $v > 0$, then $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \ni \varphi \mapsto \mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi))$ is d_1–lsc. Further, for any $\varphi \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$, there exists a sequence of smooth potentials $\varphi_j \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ such that $d_1(\varphi_j,\varphi) \to 0$ and $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi_j)) \to \mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi))$ as $j \to \infty$.*

Proof The proof follows closely the arguments of [14, Lemma 3.1]. By Lemma 6.15 and the fact that the entropy $\chi \mapsto \mathrm{Ent}(\omega_0^m, \chi)$ is lsc on the space of finite measures, with respect to the weak convergence of measures (see [11, Proposition 3.1]), it follows that the entropy $\varphi \mapsto \mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi))$ is d_1–lsc. Let $\varphi \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$. If $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi)) = \infty$ then any sequence $\varphi_j \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ such that $d_1(\varphi_j,\varphi) \to 0$ satisfies $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi)) \to \infty$ as $j \to \infty$. We suppose $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi)) < \infty$ and we put $g := \mathrm{MA}_v(\varphi)/\omega_0^{[m]} \geqslant 0$, the density function of the measure $\mathrm{MA}_v(\varphi)$. From the proof of [14, Lemma 3.1], there exist a sequence of positive functions $g_j \in C^\infty_{\mathbb{T}}(X)$ such that $\|g - g_j\|_{L^1} \to 0$ and

$$\int_X g_j \log g_j \omega_0^{[m]} \to \mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi)).$$

Using [45, Proposition 3.7], we can find a smooth potential $\varphi_j \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ (which is unique up to adding a constant) such that $\mathrm{MA}_v(\varphi_j) = \left(\int_X v(\mu_0)\omega_0^m / \int_X g_j \omega_0^m\right) g_j \omega_0^{[m]}$. By [45, Lemma 2.16], up to passing to a subsequence of φ_j, there exists a $\psi \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ such that $d_1(\psi,\varphi_j) \to 0$. Lemma 6.15 together with $\|g - g_j\|_{L^1} \to 0$ gives

$$\mathrm{MA}_v(\psi) = \lim_{j\to\infty} \mathrm{MA}_v(\varphi_j) = \mathrm{MA}_v(\varphi).$$

It follows that $\varphi = \psi$ (up to a constant) by [15, Theorem 2.18]. Thus, $d_1(\varphi,\varphi_j) \to 0$ and $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi_j)) \to \mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi))$ as $j \to \infty$. □

Now we are in position to prove Theorem 6.1.

Proof of Theorem 6.1 By Lemmas 6.10 and 6.12, the functionals $\boldsymbol{I}_w$ and $\boldsymbol{I}_v^{\rho_{\omega_0}}$ extend as continuous functionals on $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$. On the other hand, the entropy $\varphi \mapsto \mathrm{Ent}(\omega_0^m, \mathrm{MA}_v(\varphi))$ is d_1–lsc by Lemma 6.16. Thus, the weighted Chen–Tian decomposition (50) gives rise to an extension of the (v,w)–Mabuchi energy to a d_1–lsc functional $\boldsymbol{M}_{v,w} \colon \mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \to \mathbb{R} \cup \{\infty\}$. Notice that (using the continuity of $\boldsymbol{I}_w$ and $\boldsymbol{I}_v^{\rho_{\omega_0}}$) the restriction of $\boldsymbol{M}_{v,w} \colon \mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \to \mathbb{R} \cup \{\infty\}$ on the subspace $\mathcal{K}^{1,\bar{1}}_{\mathbb{T}}(X,\omega_0)$ is equal to the weighted (v,w)–Mabuchi energy on that space defined in [56, Corollary 3]. By Lemma 6.16, for $\varphi \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$, we can find a sequence $\varphi_j \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ such that $d_1(\varphi_j,\varphi) \to 0$ and

$$\lim_{j\to\infty} \boldsymbol{M}_{v,w}(\varphi_j) = \boldsymbol{M}_{v,w}(\varphi).$$

It follows that the extension $\boldsymbol{M}_{v,w}\colon \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}\cup\{\infty\}$ using (50) is the largest d_1–lsc extension of $\boldsymbol{M}_{v,w}\colon \mathcal{K}_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}$.

We now show that $t\mapsto \boldsymbol{M}_{v,w}(\varphi_t)$ for $t\in[0,1]$ is convex and continuous along the finite-energy geodesics $\varphi_t\in\mathcal{E}_{\mathbb{T}}(X,\omega_0)$. We will follow closely the arguments of [14, Theorem 4.7]. Let $\varphi_t\in\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ for $t\in[0,1]$ be a finite-energy geodesic. Suppose that $t_0,t_1\in[0,1]$ with $t_0\leqslant t_1$. Using Lemma 6.16, we can find sequences $\varphi^j_{t_0},\varphi^j_{t_1}\in\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ such that $d_1(\varphi^j_{t_0},\varphi_{t_0})\to 0$ and $d_1(\varphi^j_{t_1},\varphi_{t_1})\to 0$, and

$$\lim_{j\to\infty}\boldsymbol{M}_{v,w}(\varphi^j_{t_0})=\boldsymbol{M}_{v,w}(\varphi_{t_0})\quad\text{and}\quad \lim_{j\to\infty}\boldsymbol{M}_{v,w}(\varphi^j_{t_1})=\boldsymbol{M}_{v,w}(\varphi_{t_1}).$$

Let $t\mapsto\varphi^j_t\in\mathcal{K}^{1,\bar{1}}_{\mathbb{T}}(X,\omega_0)$ for $t\in[t_0,t_1]$ be the $C^{1,\bar{1}}$–weak geodesic segment connecting $\varphi^j_{t_0}$ and $\varphi^j_{t_1}$. By [56, Theorem 5], the function $[t_0,t_1]\ni t\mapsto\boldsymbol{M}_{v,w}(\varphi^j_t)$ is convex. Since $\boldsymbol{M}_{v,w}\colon\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}\cup\{\infty\}$ is d_1 lsc,

$$\begin{aligned}\boldsymbol{M}_{v,w}(\varphi_t)&\leqslant\liminf_{j\to\infty}\boldsymbol{M}_{v,w}(\varphi^j_t)\\&\leqslant\left(\frac{t-t_0}{t_1-t_0}\right)\lim_{j\to\infty}\boldsymbol{M}_{v,w}(\varphi^j_{t_0})+\left(\frac{t_1-t}{t_1-t_0}\right)\lim_{j\to\infty}\boldsymbol{M}_{v,w}(\varphi^j_{t_1})\\&\leqslant\left(\frac{t-t_0}{t_1-t_0}\right)\boldsymbol{M}_{v,w}(\varphi_{t_0})+\left(\frac{t_1-t}{t_1-t_0}\right)\lim_{j\to\infty}\boldsymbol{M}_{v,w}(\varphi_{t_1}),\end{aligned}$$

where the second inequality uses the convexity of $t\mapsto\boldsymbol{M}_{v,w}(\varphi^j_t)$. Thus, $t\mapsto\boldsymbol{M}_{v,w}(\varphi_t)$ is convex and continuous up to the boundary of $[t_0,t_1]$ since it is d_1–lsc.

It remains to show that $\boldsymbol{M}_{v,w}\colon\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}\cup\{\infty\}$ is linear and continuous in v and w. For smooth potentials $\varphi\in\mathcal{K}_{\mathbb{T}}(X,\omega_0)$,

$$\text{(58)}\quad \mathrm{Ent}(\omega_0^{[m]},\mathrm{MA}_v(\varphi))-\int_X\log(v(\mu_0))v(\mu_0)\omega_0^{[m]}=\int_X\log\left(\frac{\mathrm{MA}(\varphi)}{\omega_0^m}\right)\mathrm{MA}_v(\varphi),$$

which is manifestly linear in v. For $\varphi\in\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ the above expression is still linear in v by Proposition 6.13. Substituting back in (50), and using Lemmas 6.10 and 6.12, it follows that $\boldsymbol{M}_{v,w}\colon\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}\cup\{\infty\}$ is linear in v and w. From these two lemmas we know that $\boldsymbol{I}^{\rho}_v\colon\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}$ and $\boldsymbol{I}_w\colon\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}$ are uniformly continuous in v and w. For the remaining entropy part, we notice that, if $\varphi\in\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$, $v,v'\in C^\infty(\Delta)$ and $f\in C^0(X)$, then

$$\left|\int_X f\,\mathrm{MA}_v(\varphi)-\int_X f\,\mathrm{MA}_{v'}(\varphi)\right|\leqslant\|v-v'\|_{C^0(\Delta)}\int_X|f|\,\mathrm{MA}(\varphi),$$

which can be obtained again by approximating φ with a monotone sequence of smooth relative potentials and Proposition 6.13. So $C^\infty(\Delta)\times\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\ni(v,\varphi)\mapsto\mathrm{MA}_v(\varphi)$ is uniformly continuous with respect to v for the weak topology on the space of measures. Since the entropy $\chi \mapsto \mathrm{Ent}(\omega_0^m, \chi)$ is lsc on the space of finite measures with respect to the weak convergence of measures [11, Proposition 3.1], the term $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi))$ is lsc with respect to v. The linearity with respect to v in the right side of (58) shows that $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi))$ is in fact continuous with respect to v. □

We derive the following weighted version of the key compactness result from [12; 13]:

Theorem 6.17 *Any sequence $\varphi_j \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ such that*

$$d_1(0, \varphi_j) \leqslant C \quad \text{and} \quad \boldsymbol{M}_{v,w}(\varphi_j) \leqslant C$$

admits a d_1–convergent subsequence.

Proof From (50) and Lemmas 6.9 and 6.12, $\mathrm{Ent}(\omega_0^{[m]}, \mathrm{MA}_v(\varphi))$ is uniformly bounded under the hypotheses. We conclude using [45, Lemma 2.16]. □

7 Regularity of the weak minimizers of the weighted Mabuchi energy

In this section, we establish the regularity of the weak minimizers of $\boldsymbol{M}_{v,w}$.

Theorem 7.1 *Suppose $\mathbb{T} \subset \mathrm{Aut}_r(X)$ is a maximal torus and $(X, \alpha, \mathbb{T})$ admits a (v,w)–cscK metric ω with $w = \ell^{\mathrm{ext}}_{v,w_0} w_0$, where $v, w_0 > 0$ are two positive smooth weight functions on Δ. If $\psi \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$ is a minimizer of the extended (v,w)–Mabuchi energy $\boldsymbol{M}_{v,w} \colon \mathcal{E}^1_{\mathbb{T}}(X, \omega_0) \to \mathbb{R} \cup \{\infty\}$, then $\psi \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ is a smooth potential.*

The proof of this result, which is an adaptation of the arguments in [14], will occupy the reminder of the section.

Definition 7.2 Let $v(\mu) > 0$ and $w(\mu)$ be smooth weight functions on Δ and $\rho > 0$ a $\mathbb{T}$–invariant Kähler form on X. We let

$$\mathcal{M}_{v,w} := \bigl\{\psi \in \mathcal{E}^1_{\mathbb{T}}(X, \omega_0) \cap \boldsymbol{I}^{-1}(0) \mid \boldsymbol{M}_{v,w}(\psi) = \inf_{\varphi \in \mathcal{E}^1_{\mathbb{T}}} \boldsymbol{M}_{v,w}(\varphi)\bigr\} \tag{59}$$

and $\boldsymbol{M}^\rho_{v,w} := \boldsymbol{M}_{v,w} + \boldsymbol{I}^\rho$, where $\boldsymbol{I}^\rho$ is introduced via Lemma 6.12 and $v = 1$.

By [29, Lemma 5.2] and Theorem 6.1, the set $\mathcal{M}_{v,w}$ (when nonempty) is totally geodesic with respect to the finite-energy geodesics of $\mathcal{E}^1_{\mathbb{T}}(X, \omega_0)$. Furthermore, if there exists a $\psi_\rho \in \mathcal{M}_{v,w}$ such that $\boldsymbol{I}^\rho(\psi_\rho) = \inf_{\psi \in \mathcal{M}_{v,w}} \boldsymbol{I}^\rho(\psi)$, then ψ_ρ is unique by the strict convexity of $\boldsymbol{I}^\rho$ established in [14, Proposition 4.5]. Furthermore, by Theorem 6.1, the functional $\boldsymbol{M}^\rho_{v,w} : \mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \to \mathbb{R}\cup\{\infty\}$ will also be strictly convex along finite-energy geodesics, showing the uniqueness of an element $\psi \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$ such that $\boldsymbol{M}^\rho_{v,w}(\psi) = \inf_{\varphi\in\mathcal{E}^1_{\mathbb{T}}} \boldsymbol{M}^\rho_{v,w}(\varphi)$ (assuming that such minimizer ψ exists).

We then have a weighted version of the continuity method of [14, Proposition 3.1]:

Proposition 7.3 *Let $v > 0$ and w be smooth weight functions on Δ. Suppose that $\mathcal{M}_{v,w}$ is nonempty and $\varphi \in \mathcal{K}_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$. Then, for any $\lambda > 0$, there exists a unique minimizer $\psi_\lambda \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$ of $\boldsymbol{M}^{\lambda\omega_\varphi}_{v,w} := \boldsymbol{M}_{v,w} + \boldsymbol{I}^{\lambda\omega_\varphi}$. The curve $[0,\infty) \ni \lambda \mapsto \psi_\lambda \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$ is d_1–continuous and d_1–bounded, and $\psi_0 := \lim_{\lambda\to 0}\psi_\lambda$ is the unique minimizer of $\boldsymbol{I}^{\omega_\varphi}$ on $\mathcal{M}_{v,w}$. Furthermore, for any $\psi \in \mathcal{M}_{v,w}$ and $\lambda > 0$,*

$$I(\varphi,\psi_\lambda) \leqslant m(m+1) I(\varphi,\psi), \tag{60}$$

where $I(\varphi,\psi) := \int_X (\varphi-\psi)(\omega^m_\psi - \omega^m_\varphi)$.

Proof The proof is a straightforward adaptation of that of [14, Proposition 3.1]. □

We next need a weighted analogue of [14, Lemma 3.3]:

Lemma 7.4 *Let $v > 0$ and w be smooth weight functions on Δ, and $\rho > 0$ a smooth $\mathbb{T}$–invariant Kähler form on X. Let $\varphi_0 \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ and $\varphi_1 \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$, and $[0,1] \ni t \mapsto \varphi_t \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ be a finite-energy geodesic connecting φ_0 and φ_1. Then*

$$\lim_{t\to 0^+} \frac{\boldsymbol{M}^\rho_{v,w}(\varphi_t) - \boldsymbol{M}^\rho_{v,w}(\varphi_0)}{t} \geqslant \int_X (w(\mu_{\varphi_0}) - \mathrm{Scal}_v(\varphi_0))\dot\varphi_0 \omega^{[m]}_{\varphi_0} + \int_X \dot\varphi_0 \rho \wedge \omega^{[m-1]}_{\varphi_0},$$

where $\boldsymbol{M}^\rho_{v,w} := \boldsymbol{M}_{v,w} + \boldsymbol{I}^\rho$.

Proof By Theorem 6.1 and the fact that $\boldsymbol{I}^\rho$ is d_1–continuous (see [14] or Lemma 6.12), for any $t \in [0,1]$ there exists a sequence $(\varphi^k_t)_k \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ such that

$$\lim_{k\to\infty} d_1(\varphi^k_t, \varphi_t) = 0 \quad \text{and} \quad \boldsymbol{M}^\rho_{v,w}(\varphi^k_t) \to \boldsymbol{M}^\rho_{v,w}(\varphi_t).$$

We let $[0,t] \ni s \mapsto \psi^k_s$ be the weak $C^{1,\bar 1}$–geodesic joining $\varphi^k_0 = \varphi_0$ with φ^k_t. By the proof of [56, Corollary 1],

$$\lim_{t\to 0^+} \frac{\boldsymbol{M}^\rho_{v,w}(\varphi^k_t) - \boldsymbol{M}^\rho_{v,w}(\varphi_0)}{t} \geqslant \int_X (w(\mu_{\varphi_0}) - \mathrm{Scal}_v(\varphi_0))\dot\psi^k_0 \omega^{[m]}_{\varphi_0} + \int_X \dot\psi^k_0 \rho\wedge\omega^{[m-1]}_{\varphi_0}.$$

According to [14, Lemma 3.4], we can use the dominated convergence theorem on the right side of the above inequality to conclude. □

The last step is to establish a weighted version of [14, Proposition 3.2.]:

Proposition 7.5 *Suppose $\mathbb{T} \subset \mathrm{Aut}_r(X)$ is a maximal torus, and let $v(\mu), w_0(\mu) > 0$ and $w = \ell^{\mathrm{ext}}_{v,w_0} w_0$. Suppose that $\varphi^* \in \mathcal{K}_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$ is a (v,w)–cscK potential. Then, for any fixed Kähler form ω_φ with $\varphi \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$, there exists a $\sigma \in \mathbb{G} := \mathbb{T}^{\mathbb{C}}$ such that*

$$\inf_{\psi \in \mathcal{M}_{v,w}} \boldsymbol{I}^{\omega_\varphi}(\psi) = \boldsymbol{I}^{\omega_\varphi}(\sigma[\varphi^*]).$$

Proof As $\mathbb{G}$ is reductive, there exists a unique $\sigma \in \mathbb{G}$ such that

$$\boldsymbol{I}^{\omega_\varphi}(\sigma[\varphi^*]) = \inf_{\tau \in \mathbb{G}} \boldsymbol{I}^{\omega_\varphi}(\tau[\varphi^*]) \tag{61}$$

(see eg [29, Section 6] or [56, Lemma 11]), where, we recall, the $\mathbb{G}$ action on potentials is introduced via the slice $\boldsymbol{I}^{-1}(0)$. Let $\varphi_0 := \sigma[\varphi^*] \in \mathcal{K}_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$, and $\psi_0 \in \mathcal{M}_{v,w}$ be the unique minimizer of $\boldsymbol{I}^{\omega_\varphi}$. We want to show that $\varphi_0 = \psi_0$.

For $\lambda > 0$ let the unique minimizer of $\boldsymbol{M}^{\lambda\omega_\varphi}_{v,w} = \boldsymbol{M}_{v,w} + \lambda \boldsymbol{I}^{\omega_\varphi}$ on $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$ be ψ_λ, as given by Proposition 7.3. By this proposition, $\lim_{\lambda\to 0} d_1(\psi_\lambda,\psi_0) = 0$. We denote by $\boldsymbol{V}_\lambda$ and $\boldsymbol{W}$ the differentials of $\boldsymbol{M}^{\lambda\omega_\varphi}_{v,w}$ and $\boldsymbol{I}^{\omega_\varphi}$, respectively, viewed as 1–forms on the Fréchet space $\mathcal{K}(X,\omega_0)$. We thus have, for all $\psi \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ and for all $\dot\psi \in C^\infty_{\mathbb{T}}(X)$,

$$\begin{aligned}(\boldsymbol{V}_0)_\psi(\dot\psi) &= -\int_X (\mathrm{Scal}_v(\omega_\psi) - w(\mu_\psi))\dot\psi\,\omega_\psi^{[m]},\\ \boldsymbol{W}_\psi(\dot\psi) &= \int_X \dot\psi\,\omega_\varphi \wedge \omega_\psi^{[m-1]},\\ (\boldsymbol{V}_\lambda)_\psi(\dot\psi) &= (\boldsymbol{V}_0)_\psi(\dot\psi) + \lambda \boldsymbol{W}_\psi(\dot\psi).\end{aligned} \tag{62}$$

Recall that the Mabuchi connection $\mathcal{D}$ on the Fréchet space $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$ is introduced by

$$(\mathcal{D}_{\dot\varphi_t}\dot\psi_t)_{\varphi_t} := \ddot\psi_t - \langle d\dot\psi_t, d\dot\varphi_t\rangle_{\omega_{\varphi_t}},$$

where φ_t and ψ_t are smooth paths in $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$. Using [55, Lemma B.1], we compute the covariant derivative of $\boldsymbol{V}_0$ with respect to the Mabuchi connection to be

$$\begin{aligned}((\mathcal{D}_{\dot\psi_2}\boldsymbol{V}_0)(\dot\psi_1))_\psi = \int_X \big[&2v(\mu_\psi)((\nabla^{\omega_\psi} d\dot\psi_1)^-, (\nabla^{\omega_\psi} d\dot\psi_2)^-)_{\omega_\psi}\\ &+ (\mathrm{Scal}_v(\omega_\psi) - w(\mu_\psi))(d\dot\psi_1, d\dot\psi_2)_{\omega_\psi}\big]\omega_\psi^{[m]},\end{aligned}$$

where $(\nabla^{\omega_\psi} d\dot\psi)^-$ denotes the $(2,0)+(0,2)$ part of the Hessian of $\dot\psi$ with respect to the Levi-Civita connection ∇^{ω_ψ} of ω_ψ. Taking $\psi = \varphi_0$ to be the (v,w)–cscK potential,

$$\begin{aligned}((\mathcal{D}_{\dot\psi_2}\boldsymbol{V}_0)(\dot\psi_1))_{\varphi_0} &= 2\int_X ((\nabla^{\omega_{\varphi_0}} d\dot\psi_1)^-, (\nabla^{\omega_{\varphi_0}} d\dot\psi_2)^-)_{\omega_{\varphi_0}} v(\mu_{\varphi_0})\omega_{\varphi_0}^{[m]}\\ &= 2\int_X \mathbb{L}_{\omega_{\varphi_0},v}(\dot\psi_1)\dot\psi_2\,\omega_{\varphi_0}^{[m]} = 2\int_X \mathbb{L}_{\omega_{\varphi_0},v}(\dot\psi_2)\dot\psi_1\,\omega_{\varphi_0}^{[m]},\end{aligned}$$

where the operator $\mathbb{L}_{\omega_\psi,v}(\dot\psi) := \delta_{\omega_\psi}\delta_{\omega_\psi}(v(\mu_\psi)(\nabla^{\omega_\psi} d\dot\psi)^-)$ is a fourth-order elliptic self-adjoint operator on (X,ω_ψ), with kernel given by the space of Killing potentials in $C^\infty_{\mathbb{T}}(X)$; see Appendix A.

As φ_0 is a (v,w)–cscK potential which satisfies (61), we have by [56, Lemma 10] that $\boldsymbol{W}_{\varphi_0}(\dot\psi) = 0$ for any $\mathbb{T}$–invariant Killing potential $\dot\psi$ with respect to ω_{φ_0}. It follows that we can solve the linear equation (for a function $\dot\psi \in C^\infty_{\mathbb{T}}(X)$)

$$\mathbb{L}_{\omega_{\varphi_0},v}(\dot\psi) = \frac{\omega_\varphi \wedge \omega_{\varphi_0}^{[m-1]}}{\omega_{\varphi_0}^{[m]}},$$

as the right side is L^2–orthogonal (with respect to the measure $\omega_{\varphi_0}^{[m]}$) to the kernel of $\mathbb{L}_{\omega_{\varphi_0},v}$. Equivalently, there exists a $\dot\psi_0 \in C^\infty_{\mathbb{T}}(X)$ such that we have equality of 1–forms on $\mathcal{K}_{\mathbb{T}}(X,\omega_0)$:

$$(\mathcal{D}_{\dot\psi_0}\boldsymbol{V}_0)_{\varphi_0} = -\boldsymbol{W}_{\varphi_0}. \tag{63}$$

Let $\lambda \to \dot\phi_\lambda \in C^\infty_{\mathbb{T}}(X)$ be a smooth curve in the tangent space to $(\varphi_0+\lambda\dot\psi_0) \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$, defined for $\lambda > 0$ small enough. We compute

$$\begin{aligned}\frac{d}{d\lambda}\Big|_{\lambda=0}(\boldsymbol{V}_\lambda)_{\varphi_0+\lambda\dot\psi_0}(\dot\phi_\lambda) &= \boldsymbol{W}_{\varphi_0}(\dot\phi_0) + ((\mathcal{D}_{\dot\psi_0}\boldsymbol{V}_0)(\dot\phi_0))_{\varphi_0} + (\boldsymbol{V}_0)_{\varphi_0}\Big(\frac{d}{d\lambda}\Big|_{\lambda=0}\dot\phi_\lambda\Big)\\ &= 0,\end{aligned} \tag{64}$$

where we have used (63) and that $(\boldsymbol{V}_0)_{\varphi_0} = 0$ since φ_0 is a (v,w)–cscK potential; see (62). On the other hand, letting

$$f_\lambda := -\mathrm{Scal}_v(\omega_{\varphi_0+\lambda\dot\psi_0}) + w(\mu_{\varphi_0+\lambda\dot\psi_0}) + \langle \omega_{\varphi_0+\lambda\dot\psi_0}, \omega_\varphi\rangle_{\omega_\varphi},$$

it follows from (62) that, for any $\dot\phi \in C^\infty_{\mathbb{T}}(X)$,

$$(\boldsymbol{V}_\lambda)_{\varphi_0+\lambda\dot\psi_0}(\dot\phi) = \int_X \dot\phi f_\lambda \omega^{[m]}_{\varphi_0+\lambda\dot\psi_0}.$$

Thus (64) implies that $f_\lambda = O(\lambda^2)$ and

$$|(\boldsymbol{V}_\lambda)_{\varphi_0+\lambda\dot\psi_0}(\dot\phi)| \leqslant C\lambda^2 \sup_X |\dot\phi|.$$

Let $\psi_\lambda(t) \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ be a finite-energy geodesic connecting $\psi_\lambda(0) := \psi_\lambda \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ with $\psi_\lambda(1) := \varphi_0 + \lambda\dot\psi_0 \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$ for $\lambda > 0$ small enough. By Lemma 7.4,

$$\frac{d}{dt}\Big|_{t=1^-}\boldsymbol{M}^{\lambda\omega_\varphi}_{v,w}(\psi_\lambda(t)) \leqslant \int_X \dot\psi_\lambda(1) f_\lambda \omega^{[m]}_{\varphi_0+\lambda\dot\psi_0}.$$

By Proposition 7.3, $d_1(0,\psi_\lambda(0))$ is uniformly bounded. Also, $d_1(0,\psi_\lambda(1))$ is uniformly bounded for λ small enough since $\psi_\lambda(1) := \varphi_0 + \lambda\dot\psi_0 \in \mathcal{K}_{\mathbb{T}}(X,\omega_0)$. We thus have that both $d_1(0,\psi_\lambda(0))$ and $d_1(0,\psi_\lambda(1))$ are uniformly bounded and, by [14, Lemma 3.4(ii)], we get

$$\int_X |\dot\psi_\lambda(1)|\omega^{[m]}_{\varphi_0+\lambda\dot\psi_0} = d_1(\psi_\lambda(0),\psi_\lambda(0)) \leqslant d_1(0,\psi_\lambda(0)) + d_1(0,\psi_\lambda(1)) \leqslant C.$$

From $f_\lambda = O(\lambda^2)$, we obtain

$$\frac{d}{dt}\Big|_{t=1^-} \boldsymbol{M}^{\lambda\omega_\varphi}_{v,w}(\psi_\lambda(t)) \leqslant O(\lambda^2).$$

As the unique minimizer of the strictly convex functional $\boldsymbol{M}^{\lambda\omega_\varphi}_{v,w}$ on $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\cap \boldsymbol{I}^{-1}(0)$ is $\psi_\lambda(0)=\psi_\lambda$,

$$\frac{d}{dt}\Big|_{t=1^-} \boldsymbol{M}^{\lambda\omega_\varphi}_{v,w}(\psi_\lambda(t)) \geqslant \frac{d}{dt}\Big|_{t=0^+} \boldsymbol{M}^{\lambda\omega_\varphi}_{v,w}(\psi_\lambda(t)) \geqslant 0.$$

Using that the functions $t\mapsto \boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(t))$ and $t\mapsto \boldsymbol{M}_{v,w}(\psi_\lambda(t))$ are both convex (this follows from [14, Proposition 4.5] and Theorem 6.1),

$$\begin{aligned} 0 \leqslant \lambda\Big(\frac{d}{dt}\Big|_{t=1^-} - \frac{d}{dt}\Big|_{t=0^+}\Big)\boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(t)) &\leqslant \Big(\frac{d}{dt}\Big|_{t=1^-} - \frac{d}{dt}\Big|_{t=0^+}\Big)\boldsymbol{M}^{\lambda\omega_\varphi}_{v,w}(\psi_\lambda(t)) \\ &\leqslant O(\lambda^2). \end{aligned}$$

By the convexity of $t\mapsto \boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(t))$, the last estimate also gives

$$\begin{aligned} 0 &\leqslant t\boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(1)) + (1-t)\boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(0)) - \boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(t)) \\ &= t(1-t)\bigg(\frac{\boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(1)) - \boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(t))}{1-t}\bigg) - t(1-t)\bigg(\frac{-\boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(0)) + \boldsymbol{I}^{\omega_\varphi}(\psi_\lambda(t))}{t}\bigg) \\ &\leqslant t(1-t)O(\lambda). \end{aligned}$$

Letting $\lambda\to 0$ and using the endpoint stability of the finite-energy geodesic segments (see [14, Proposition 4.3]) together with the d_1–continuity of $\boldsymbol{I}^{\omega_\varphi}$ [14, Proposition 4.4], $t\mapsto \boldsymbol{I}^{\omega_\varphi}(\psi(t))$ is linear along the finite-energy geodesic $\psi(t)=\lim_{\lambda\to 0^+}\psi_\lambda(t)$ connecting $\psi_0(0)=\psi_0$ and $\psi_0(1)=\varphi_0$. The strict convexity of $\boldsymbol{I}^{\omega_\varphi}$ along finite-energy geodesics [14, Proposition 4.5] then yields $\psi_0=\varphi_0=\sigma[\varphi^*]$. □

Now we are in position to prove Theorem 7.1 by the arguments in [14, Theorem 1.4].

Proof of Theorem 7.1 Without loss of generality, we can assume that the (v,w)–extremal metric $\omega^*=\omega_0$ is the initial metric, and we suppose $\psi_0\in\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\cap \boldsymbol{I}^{-1}(0)$ is a weak minimizer of $\boldsymbol{M}_{v,w}:\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}\cup\{\infty\}$. We want to show that $\psi_0=\sigma[0]$ for some $\sigma\in\mathbb{G}=\mathbb{T}^{\mathbb{C}}$. It is well known (see [28] or Corollary 6.11) that there exists a sequence $\varphi_j\in\mathcal{K}_{\mathbb{T}}(X,\omega_0)\cap \boldsymbol{I}^{-1}(0)$ such that $d_1(\varphi_j,\psi_0)\to 0$. We set $\rho_j=\omega_0+dd^c\varphi_j$, which is a $\mathbb{T}$–invariant Kähler form.

Since ω_0 is a (v,w)–extremal metric, $\mathcal{M}_{v,w}$ is nonempty. By Proposition 7.3, the functional $\boldsymbol{M}^{\lambda\rho_j}_{v,w} = \boldsymbol{M}_{v,w} + \lambda \boldsymbol{I}^{\rho_j}$ has a unique minimizer $\psi_{j,\lambda} \in \mathcal{E}^1_{\mathbb{T}}(X,\omega_0) \cap \boldsymbol{I}^{-1}(0)$ such that

$$I(\varphi_j, \psi_{j,\lambda}) \leqslant m(m+1) I(\varphi_j, \psi_0).$$

By the quasitriangle identity [14, (2.16)],

$$I(\psi_0, \psi_{j,\lambda}) \leqslant C(I(\psi_0,\varphi_j) + I(\varphi_j,\psi_{j,\lambda})) \leqslant C(m^2+m+1) I(\varphi_j,\psi_0), \tag{65}$$

where $C > 0$ is a uniform constant depending only on m.

Let $j > 0$ be fixed. According to Proposition 7.3, $\psi_{j,0} := \lim_{\lambda\to 0} \psi_{j,\lambda}$ is the unique minimizer of $\boldsymbol{I}^{\lambda\rho_j}$ on $\mathcal{M}_{v,w}$, whereas Proposition 7.5 yields that there exists a $\sigma_j \in \mathbb{G}$ such that $\psi_{j,0} = \sigma_j[0]$. Letting $\lambda \to 0^+$ in (65) (and using the d_1–continuity of I; see eg [13] or Lemma 6.10),

$$I(\psi_0, \sigma_j[0]) \leqslant C(m^2+m+1) I(\varphi_j,\psi_0).$$

When $j\to\infty$ (using $d_1(\varphi_j,\psi_0)\to 0$), we get $I(\psi_0,\sigma_j[0])\to 0$. By [12, Proposition 2.3; 27, Proposition 5.9], the latter limit is equivalent to $d_1(\sigma_j[0],\psi_0) \to 0$. Using [14, Lemma 3.7], there exists a $\sigma \in \mathbb{G}$ such that $\sigma[0] = \psi_0$. □

Remark 7.6 The arguments in the proofs of Proposition 7.5 and Theorem 7.1 extend if we remove the maximality assumption for $\mathbb{T} \subset \mathrm{Aut}_r(X)$, and replace $\mathbb{G} = \mathbb{T}^{\mathbb{C}}$ with the connected component of the identity $\widehat{\mathbb{G}} = \mathrm{Aut}^{\mathbb{T}}_r(X)$ of the centralizer of $\mathbb{T}$ in $\mathrm{Aut}_r(X)$. The key points are that $\widehat{\mathbb{G}}$ is reductive (see Proposition 1.4) and $\widehat{\mathbb{G}}$ acts transitively on the space of $\mathbb{T}$–invariant (v,w_0)–extremal Kähler metrics (see Theorem 1.5).

Proof of Theorem 1 We apply the coercivity principle of [29]; see Theorem 3.6. By Theorem 6.1, the extension of the weighted Mabuchi energy $\boldsymbol{M}_{v,w}$ to the space $\mathcal{E}^1_{\mathbb{T}}(X,\omega_0)$ satisfies the hypotheses of Theorem 3.6 (the invariance of $\boldsymbol{M}_{v,w}$ under the action of $\mathbb{G} = \mathbb{T}^{\mathbb{C}}$ is equivalent to the necessary condition (3) for the existence of a (v,w)–cscK metric). We thus need to ensure that $\boldsymbol{M}_{v,w}$ further satisfies properties (i)–(iv) of Theorem 3.6. Theorem 6.1 also yields the convexity property (i), whereas the regularity property (ii) is established in Theorem 7.1. This last result also yields the uniqueness property (iii) via Theorem 1.5. Finally, the compactness property (iv) is established in Theorem 6.17. □

Remark 7.7 By virtue of Theorem 1.5 and Remark 7.6, the conclusion of Theorem 1 holds true if one drops the assumption that $\mathbb{T} \subset \mathrm{Aut}_r(X)$ is a maximal torus, but instead of $\mathbb{T}^{\mathbb{C}}$ one considers the larger reductive group $\widehat{\mathbb{G}} = \mathrm{Aut}^{\mathbb{T}}_r(X)$; see Proposition 1.4.

8 Proofs of Theorems 2 and 3

Proof of Theorem 2 The implication (ii) $\Longrightarrow$ (i) follows from Lemma 5.9, whereas (ii) $\Longrightarrow$ (iii) is established in Theorem 1. We shall prove (iii) $\Longrightarrow$ (ii) and (i) $\Longrightarrow$ (ii). The arguments are very similar to the ones in the proof of [52, Theorem 1], where the case when $(X, \mathbb{T})$ is toric is studied. The main idea is to show that, on a semisimple principal $(X, \mathbb{T})$–fibration, the continuity path used by Chen and Cheng [23] in the cscK case and its modification by He [47] to the extremal case can be adapted to bundle-compatible construction. We sketch the proof below for the reader's convenience.

(iii) $\Longrightarrow$ (ii) We shall work on Y. Let $\tilde{\omega}_0$ be a bundle-compatible Kähler metric on Y, corresponding to a $\mathbb{T}_X$–invariant Kähler metric ω_0 on X. By Lemma 5.14, $\tilde{\omega}_0$ is invariant under a maximal torus $\mathbb{K}_Y \subset \mathrm{Aut}_r(Y)$ (containing $\mathbb{T}_Y$), and, by this lemma and Lemma 5.10, the extremal affine-linear function corresponding to $\mathbb{K}_Y$ is the pullback to the vector space $\mathfrak{k}_Y^* = (\mathrm{Lie}(\mathbb{K}_Y))^*$ of the extremal affine-linear function $\ell^{\mathrm{ext}}(\mu)$ on $\mathfrak{t}$ defined in Theorem 2(ii). Furthermore, by Lemma 5.10, the restriction of $\boldsymbol{M}^Y_{1,\ell^{\mathrm{ext}}}$ to the subspace $\mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0)$ (see Corollary 5.6 and Lemma 5.14) is a positive multiple of $\boldsymbol{M}^X_{p,\tilde{w}}$, where the weights are those defined in Theorem 2(ii). In this setup, the main ingredients of the proof are as follows.

Step 1 Following [23; 46; 47], one considers the continuity path $\varphi_t \in \mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0)$ determined by the solution of the PDE

$$t(\mathrm{Scal}(\tilde{\omega}_{\varphi_t}) - \ell^{\mathrm{ext}}(\mu_{\tilde{\omega}_{\varphi_t}})) = (1-t)(\mathrm{tr}_{\tilde{\omega}_{\varphi_t}}(\tilde{\rho}) - (n+m)) \quad \text{for } t \in (0,1), \tag{66}$$

where $\tilde{\rho}$ is a suitable (fixed) $\mathbb{K}_Y$–invariant Kähler metric on Y in the class $[\tilde{\omega}_0]$. By [23; 47] there exists $\tilde{\rho} \in [\tilde{\omega}_0]$ and a $t_0 \in (0,1)$ such that a solution φ_t of (66) exists for t in the interval $[t_0, 1)$. Furthermore, the solution $\varphi_t(y)$ is smooth as a function on $[t_0, 1) \times Y$. The main observation of [52] is that, with a suitable choice for $\tilde{\rho}$, the path (66) can in fact be reduced to a continuity path on X. To see this we observe that, by [47, Proposition 3.1], one can take $\tilde{\rho}$ in (66) to be of the form $\tilde{\rho} = \tilde{\omega}_0 + (1/r_0) dd^c f$ with r_0 large enough, where f is the smooth function on Y with zero mean with respect to $\tilde{\omega}_0$ which solves the Laplace equation

$$\Delta_{\tilde{\omega}_0} f = \mathrm{Scal}(\tilde{\omega}_0) - \ell^{\mathrm{ext}}(\mu_{\tilde{\omega}_0}).$$

By Lemmas 5.9 and A.3, $f \in C^\infty_{\mathbb{T}}(X)$, whereas by Lemma 5.5 $\tilde{\rho}$ is bundle-compatible, ie

$$\tilde{\rho} = \rho + \sum_{a=1}^{k} (\langle p_a, \mu_\rho \rangle + c_a) \pi_B^* \omega_a + \langle d\mu_\rho \wedge \theta \rangle,$$

where $\rho = \omega_0 + (1/r_0) dd^c f$ is a $\mathbb{T}$–invariant Kähler metric on X; see (24). Using Lemma 5.9 and that both $\tilde{\omega}_\varphi$ and $\tilde{\rho}$ are of the form (24), we get a path of PDEs on X of the form

$$t(\mathrm{Scal}_p(\omega_{\varphi_t}) - \tilde{w}(\mu_{\omega_{\varphi_t}})) = (1-t)H(\varphi_t) \quad \text{for } t \in (t_0, 1), \tag{67}$$

where $\varphi_t \in \mathcal{K}_{\mathbb{T}}(X, \omega_0)$ and $H(\varphi_t) := \mathrm{tr}_{\tilde{\omega}_{\varphi_t}}(\tilde{\rho}) - (n+m)$ is manifestly a second-order differential operator on X for $\varphi_t \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0)$. Then the solution φ_t for $t \in [t_0, 1)$ of (66) will actually belong to $\mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0)$. This last point is a consequence of the implicit function theorem (used in [46; 47] to establish the openness), which can be applied directly to (67); to find the linearization of (67), we use [46] that the linearization of $H(\varphi)$ on Y is the operator $\mathbb{H}^{\tilde{\rho}}_{\tilde{\omega}_\varphi, 1}$ (see Definition A.1), so that, by virtue of Lemma A.3, the linearization of $H(\varphi)$ when restricted to $\mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0)$ is given by the p–weighted operator $\mathbb{H}^{\rho}_{\omega_\varphi, p}$ introduced in Appendix A. Similar argument allows us to identify the linearization of $\mathrm{Scal}_p(\omega_\varphi)$; see also [55, Lemma B1]. We refer the reader to [52, Section 6] for further details.

Step 2 The next ingredient is a deep result from [23] with a complement in [47], showing that if $\boldsymbol{M}^Y_{1,\ell^{\mathrm{ext}}}$ is G–coercive along the continuity path φ_t with respect to a reductive subgroup $G \subset \mathrm{Aut}_r(Y)$ containing the torus generated by the extremal vector field $\xi^Y_{\mathrm{ext}} = d\ell^{\mathrm{ext}} \in \mathfrak{t}_Y$ in its centre, then there exists a subsequence of times $j \to 1$ and elements $\sigma_j \in G$ such that $\sigma_j^*(\tilde{\omega}_{\varphi_j})$ converges in $C^\infty(Y)$ to an extremal Kähler metric $\tilde{\omega}_1$. In our case, assuming (iii), we have that $\boldsymbol{M}^Y_{1,\ell^{\mathrm{ext}}}(\varphi_t) = \mathrm{Vol}(B, \omega_B) \boldsymbol{M}^X_{p,\tilde{w}}(\varphi_t)$ (see Lemma 5.10) is $\mathbb{G} = \mathbb{T}^{\mathbb{C}}_Y$–coercive (see Lemmas 6.6 and 6.4 and Proposition 3.4). We can thus find $\sigma_j \in \mathbb{T}^{\mathbb{C}}_Y$ and φ_j as above. The Kähler metrics $\sigma_j^*(\tilde{\omega}_{\varphi_j})$ are bundle-compatible in the sense of Definition 5.3, and thus are of the form

$$\sigma_j^*(\tilde{\omega}_{\varphi_j}) = \tilde{\omega}_0 + d_Y d^c_Y \sigma_j[\varphi_j] \quad \text{for } \sigma_j[\varphi_j] \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0).$$

It follows that $\tilde{\omega}_1$ is a bundle-compatible extremal Kähler metric on Y (as $\mathcal{K}_{\mathbb{T}}(X, \omega_0)$ is $C^\infty(Y)$–closed in $\mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0)$). By Lemma 5.9, the corresponding Kähler metric ω_1 on X is then $(p, \tilde{w})$–cscK.

(i) $\Longrightarrow$ (ii) The proof is very similar to the proof of (iii) $\Longrightarrow$ (ii). As in Step 1 above, we consider the continuity path (66), which defines potentials

$$\varphi_t \in \mathcal{K}_{\mathbb{T}}(X, \omega_0) \subset \mathcal{K}_{\mathbb{K}_Y}(Y, \tilde{\omega}_0) \quad \text{for } t \in [t_0, 1).$$

We can assume without loss of generality [21] that Y admits a $\mathbb{K}_Y$–invariant extremal Kähler metric in $[\tilde{\omega}_0]$, where $\mathbb{K}_Y \subset \mathrm{Aut}_r(Y)$ is the maximal torus given by Lemma 5.14.

This implies that $\boldsymbol{M}^{Y}_{1,\ell^{\mathrm{ext}}}$ is G–coercive for $G=\mathbb{K}_Y^{\mathbb{C}}$. Indeed, this can be justified, for instance, by applying Theorem 1 and Proposition 3.4 in the case $(v,w)=(1,\ell^{\mathrm{ext}})$. As in Step 2 of the proof of (iii) $\Longrightarrow$ (ii), we use [23; 47] and the G–coercivity of $\boldsymbol{M}^{Y}_{1,\ell^{\mathrm{ext}}}$ along the path in order to find a subsequence of times $j\to 1$ and elements $\sigma_j\in G$ such that $\sigma_j^*(\tilde{\omega}_{\varphi_j})$ converges in $C^\infty(Y)$ to a $\mathbb{K}_Y$–invariant extremal Kähler metric $\tilde{\omega}_1\in[\tilde{\omega}_0]$. However, unlike the proof of (iii) $\Longrightarrow$ (ii), in general $\sigma_j^*(\tilde{\omega}_{\varphi_j})$ and hence $\tilde{\omega}_1$ are not bundle-compatible, as σ_j can act nontrivially on B (see the proof of Lemma 5.14). We thus need to slightly modify the argument in order to show that $\tilde{\omega}_1$ still induces a $(p,\tilde{w})$–cscK metric on any given fibre $X_b=\pi_B^{-1}(b)\subset Y$. We denote by $\omega_j(b):=(\tilde{\omega}_{\varphi_j})|_{X_b}$ and $\bar{\omega}_j(b):=(\sigma_j^*(\tilde{\omega}_{\varphi_j}))|_{X_b}$ the induced $\mathbb{T}$–invariant metrics on X_b. As $\tilde{\omega}_{\varphi_j}$ is bundle-compatible, Lemma 5.9 yields

$$\mathrm{Scal}_p(\omega_j(b))=[p(\mu_{\tilde{\omega}_{\varphi_j}})\,\mathrm{Scal}(\tilde{\omega}_{\varphi_j})-p(\mu_{\tilde{\omega}_{\varphi_j}})q(\mu_{\tilde{\omega}_{\varphi}})]|_{X_b}.$$

Using that $\sigma_j\in\mathbb{K}_Y^{\mathbb{C}}$ sends the fibre X_b to the fibre $X_{\sigma_j(b)}$ (this follows from the construction of $\mathbb{K}_Y$ in the proof of Lemma 5.14), the above equality holds true for the metrics $\bar{\omega}_j(b)$, where in the right side we replace the metric $\tilde{\omega}_{\varphi_j}$ on Y with $\bar{\omega}_j:=\sigma_j^*(\tilde{\omega}_{\varphi_j})$. It thus follows by the smooth convergence of $\bar{\omega}_j(b)$ to $\omega_1(b)$ that

$$\begin{aligned}\mathrm{Scal}_p(\omega_1(b))&=[p(\mu_{\tilde{\omega}_1})\,\mathrm{Scal}(\tilde{\omega}_1)-p(\mu_{\tilde{\omega}_1})q(\mu_{\tilde{\omega}_1})]|_{X_b}\\&=[p(\mu_{\tilde{\omega}_1})(\ell^{\mathrm{ext}}(\mu_{\tilde{\omega}_1})-q(\mu_{\tilde{\omega}_1})]|_{X_b}=\tilde{w}(\mu_{\omega_1(b)}),\end{aligned}$$

where for the equalities on the second line we have used that the $\mathbb{K}_Y$–extremal function ℓ^{ext} is in $\mathrm{Aff}(\mathfrak{t}_X^*)$; see Lemma 5.14. Thus $\omega_1(b)$ is a $(v,\tilde{w})$–cscK metric on X. □

Proof of Theorem 3 Han and Li introduced a functional $\boldsymbol{M}_v^{\mathrm{HL}}\colon\mathcal{K}_{\mathbb{T}}(X,\omega_0)\to\mathbb{R}$ whose critical points are the v–solitons; see [45, Lemma 4.4]. A careful inspection using (50) shows that $\boldsymbol{M}_v^{\mathrm{HL}}(\omega)=\boldsymbol{M}_{v,w}(\omega)-\int_X\log(v(\mu_\omega))v(\mu_\omega)\omega^{[m]}$, where w is the weight function defined in Proposition 1. Thus, the difference of the two functionals is a constant independent of the choice of a $\mathbb{T}$–invariant Kähler metric $\omega\in 2\pi c_1(X)$; see eg [55]. Thus, by [45, Theorem 3.5] applied to $(X,2\pi c_1(X),\mathbb{T})$ (and weights $pv,\tilde{w}$), the $\mathbb{T}^{\mathbb{C}}$–coercivity of $\boldsymbol{M}^X_{pv,\tilde{w}}$ is equivalent to the existence of a vp–soliton on X. By Lemma 5.11, this implies that Y admits a (bundle-compatible) v–soliton.

By [45, Theorem 1.7], the $\mathbb{T}^{\mathbb{C}}$–coercivity of $\boldsymbol{M}^X_{pv,\tilde{w}}$ is also equivalent to the uniform vp–K–stability on $\mathbb{T}$–equivariant *special* test configurations. When $(X,\mathbb{T})$ is a toric Fano variety, the only such test configurations are the product test configurations, and thus, by [55, Proposition3], the condition is reduced to verifying (3) on X with respect to the weights $(pv,\tilde{w})$.

Therefore, in order to show the existence of a Kähler–Ricci soliton, it is sufficient to find $\xi_0 \in \mathfrak{t}$ such that (3) is satisfied for the weight functions $v(\mu) = e^{\langle \xi_0, \mu\rangle} p(\xi)$ and $\tilde{w}(\mu) = 2p(\mu)e^{\langle \xi_0,\mu\rangle}(m + \langle \xi_0, \mu\rangle + \langle d\log p, \mu\rangle)$. We detail the proof of this fact below.

Let $\omega \in 2\pi c_1(X)$ be any $\mathbb{T}$–invariant Kähler metric with canonically normalized momentum map $\mu_\omega \colon X \to \Delta$. We then consider the following p–weighted version of a functional on $\mathfrak{t}$, defined originally by Tian and Zhu [71, Lemma 2.2]:

$$\xi \mapsto \int_X e^{\langle \xi, \mu_\omega\rangle} p(\mu_\omega)\omega^{[m]} \quad \text{for } \xi \in \mathfrak{t}. \tag{68}$$

The convexity and properness of the above functional follow by the arguments in [71, Lemma 2.2], but under our toric assumption these can also be seen directly by rewriting the right side in (68) as an integral over the Delzant polytope:

$$\xi \mapsto (2\pi)^m \int_\Delta e^{\langle \xi,\mu\rangle} p(\mu)\, d\mu.$$

Properness of this functional follows by the fact that the origin is in the interior of Δ (by the canonical normalization condition of Δ; see Remark 2.3). Let $\xi_0 \in \mathfrak{t}$ be the unique critical point of (68). We have that

$$\int_X \langle \zeta, \mu_\omega\rangle e^{\langle \xi_0,\mu_\omega\rangle} p(\mu_\omega)\omega^{[m]} = 0,$$

which is precisely the condition $\mathrm{Fut}_{v,\tilde{w}} = 0$ according to Lemma B.1.

The existence of a Sasaki–Einstein structure follows similarly. By Proposition 2, Lemma 5.11 and Proposition 1, in that order, we want to find $\xi_0 \in \mathfrak{t}$ such that (3) holds true for the weights given as in Proposition 1, with $v(\mu) = p(\mu)(\langle \xi_0, \mu\rangle + a)^{-(m+n+2)}$. (This will show the existence of a pv–soliton on the toric Fano manifold $(X, \mathbb{T})$ and hence a v–soliton on Y by the general arguments evoked above.) We argue based on [61], which introduced the *volume functional* on the space of *normalized* positive affine-linear functions on Δ. Strictly speaking, the functional in [61, Section 3] is introduced on the principal $\mathbb{S}^1$–bundle N over (X, ω) (which admits a natural strictly pseudoconvex CR structure $(\mathcal{D}, J)$ coming from X), and is then defined as the Sasaki volume of a $(\mathcal{D}, J)$–compatible normalized Sasaki–Reeb vector field $\hat{\xi}$ on N; using the point of view of [7] (see in particular Lemma 1.4), the volume functional can also be written on X, noting that positive affine-linear functions $\ell_\xi = \langle \xi, \mu\rangle + a$ over Δ are in bijection with Sasaki–Reeb vector fields $\hat{\xi}$ on $(N, \mathcal{D}, J)$, and the normalization condition used in [61] is equivalent to requiring $\ell_\xi(0) = a = 1$. Specifically, in our toric weighted setting, we let

$$\xi \mapsto \int_X (\langle \xi, \mu_\omega\rangle + 1)^{-(m+n+1)} p(\mu_\omega)\omega^{[m]} = (2\pi)^m \int_\Delta (\langle \xi, \mu\rangle + 1)^{-(m+n+1)} p(\mu)\, d\mu,$$

which is defined for $\xi \in \mathfrak{t}$ such that $(\langle \xi, \mu \rangle + 1) > 0$ on Δ. The properness of the functional follows by the fact that a canonically normalized Delzant polytope of a Fano toric manifold is determined by $\Delta = \{\mu : L_j(\mu) \geq 0\}$, where the affine-linear functions $L_j(\mu)$ satisfy $L_j(0) = 1$; see eg [1, Section 7.4]. The unique critical point $\xi_0 \in \mathfrak{t}$ of the above convex functional then satisfies

$$\int_X \langle \zeta, \mu_\omega \rangle (\langle \xi_0, \mu_\omega \rangle + 1)^{-(m+n+2)} p(\mu_\omega) \omega^{[m]} = 0 \quad \text{for } \zeta \in \mathfrak{t},$$

which, by Lemma B.1, is precisely the condition (3) for the weight functions considered. This concludes the proof of Theorem 3. □

Appendix A Weighted differential operators

Let $(X, \omega, \mathbb{T})$ be as in Section 1 and $v > 0$ be a positive smooth weight function defined over the polytope Δ. We denote by ∇^ω the Levi-Civita connection of the riemannian metric g_ω, and by δ_ω the formal adjoint of ∇^ω. We define the following weighted differential operators, which are self-adjoint with respect to the volume form $v(\mu_\omega)\omega^{[m]}$ on X:

Definition A.1 The *v–weighted Laplacian* of ψ is the second-order operator acting on smooth functions, defined by

$$\Delta_{\omega,v}(\psi) = \frac{1}{v(\mu_\omega)} \delta_\omega (v(\mu_\omega) d\psi). \tag{69}$$

The *v–weighted linear Lichnerowicz* operator is the fourth-order operator given by

$$\mathbb{L}_{\omega,v}(\psi) := \frac{\delta_\omega \delta_\omega (v(\mu_\omega)(\nabla^\omega d\psi)^-)}{v(\mu_\omega)}, \tag{70}$$

where $(\nabla^\omega d\phi)^-$ stands for the $(0,2)$–symmetric tensor of type $(2,0) + (0,2)$ with respect to the complex structure of X. For any $\mathbb{T}$–invariant Kähler form ρ on X, we define the second-order operator, given by

$$\mathbb{H}^\rho_{\omega,v}(\psi) := \langle \rho, dd^c\psi \rangle_\omega + \langle d\,\mathrm{tr}_\omega(\rho), d\psi \rangle_\omega + \frac{1}{v(\mu_\omega)} \langle \rho, dv(\mu_\omega) \wedge d^c\psi \rangle_\omega, \tag{71}$$

where $\mathrm{tr}_\omega(\rho) := (\rho \wedge \omega^{[m-1]})/\omega^{[m]} = \langle \rho, \omega \rangle_\omega$. The operator $\mathbb{H}^\rho_{\omega,v}$ is a v–weighted version of the linear operator used in [46].

A straightforward computation shows:

Lemma A.2 *The v–weighted Lichnerowicz operator can be written as*

$$\mathbb{L}_{\omega,v}(\psi) = \tfrac{1}{2}(\Delta_{\omega,v})^2(\psi) + \delta_{\omega,v}\big(\rho_{\omega,v}((d^c\psi)^\sharp)\big),$$

where $\delta_{\omega,v} := (1/v(\mu_\omega))\delta_\omega v(\mu_\omega)$ is the formal adjoint of the exterior derivative d on functions with respect to the weighted volume form $v(\mu_\omega)\omega^{[m]}$,

$$\rho_{\omega,v} := \rho_\omega - \tfrac{1}{2} dd^c(\log v(\mu_\omega))$$

is the Ricci form of the weighted volume form $v(\mu_\omega)\omega^{[m]}$, and $\sharp = g_\omega^{-1}$ stands for the riemannian duality between TM and T^*M by using the Kähler metric ω.

We now specialize to the case when $(Y, \tilde{\omega}, \mathbb{T}_Y)$ is a semisimple principal $(X, \omega, \mathbb{T}_X)$–fibration over B, as in Section 5. We then denote by $\Delta^Y_{\tilde{\omega}}$, $\mathbb{L}^Y_{\tilde{\omega}}$ and $(\mathbb{H}^{\tilde{\rho}}_{\tilde{\omega}})^Y$ the corresponding unweighted operators on $(Y, \tilde{\omega})$, where the Kähler form $\tilde{\rho}$ in the definition of $\mathbb{H}^{\tilde{\rho}}_{\tilde{\omega}}$ is bundle-compatible, ie given by (24) for a $\mathbb{T}_X$–invariant Kähler form ρ on X. We further let $\Delta^{B_a}_{\omega_a}$ denote the Laplacian on (B_a, ω_a), and Δ^B_x and $\mathbb{L}^B_x$ the Laplacian and Lichnerowicz operators on B, respectively, with respect to the Kähler metric $\omega_B(x) := \sum_{a=1}^k (\langle p_a, \mu_\omega(x)\rangle + c_a)\omega_a$. We thus have:

Lemma A.3 *Let ψ be a $\mathbb{T}_Y$–invariant smooth function on Y, seen as a $\mathbb{T}_X$–invariant function on $X \times B$ via (25), and $\tilde{\omega}$ a bundle-compatible $\mathbb{T}_Y$–invariant Kähler metric on Y associated to a $\mathbb{T}_X$–invariant Kähler metric ω on X. We then have*

$$\Delta^Y_{\tilde{\omega}}\psi = \Delta^X_{\omega,p}\psi_b + \Delta^B_x\psi_x,$$

$$\mathbb{L}^Y_{\tilde{\omega}}\psi = \mathbb{L}^X_{\omega,p}\psi_b + \mathbb{L}^B_x\psi_x + \Delta^B_x(\Delta^X_{\omega,p}\psi_b)_x + \Delta^X_{\omega,v}(\Delta^B_x\psi_x)_b + \sum_{a=1}^k Q_a(x)\Delta^{B_a}_{\omega_a}\psi_x$$

and

$$(\mathbb{H}^{\tilde{\rho}}_{\tilde{\omega},1})^Y\psi = (\mathbb{H}^{\rho}_{\omega,p})^X\psi_b + \sum_{a=1}^k P_a(x)\Delta^{B_a}_{\omega_a}\psi_x,$$

where $P_a(x)$ and $Q_a(x)$ are smooth $\mathbb{T}$–invariant functions on X, and ψ_x and ψ_b are the induced smooth functions on B and X, respectively, via (25).

Proof This first two equalities are established in [6] (see the proof of Lemma 8) in the special case when $(X, \omega, \mathbb{T}_X)$ is a toric variety, whereas the third identity is proved in [52] (also in the case when $(X, \mathbb{T}_X)$ is toric). These computations extend to the general setting with no substantial additional difficulty (by using Lemma A.2 for the second identity), but we include them below for the sake of self-containedness.

In the notation of Section 5,

$$(72)\quad \Delta^Y_{\tilde{\omega}}(\psi) = \begin{cases} -(d_Y d^c_Y\psi \wedge \tilde{\omega}^{[n+m-1]})/\tilde{\omega}^{[n+m]} & \text{on } Y, \\ -(d_Y d^c_Y\psi \wedge \tilde{\omega}^{[n+m-1]} \wedge \theta^{\wedge r})/(\tilde{\omega}^{[n+m]} \wedge \theta^{\wedge r}) & \text{on } Z = X \times P, \end{cases}$$

where $\theta^{\wedge r} := \bigwedge_{i=1}^{r} \theta_i$ with respect to any lattice basis $(\xi_i)_i$ of $\Lambda \subset \mathfrak{t}$. Viewing $d^c_{X\times B}\psi$ as a 1–form on Z, it admits the decomposition, with respect to (23),

$$d^c_{X\times B}\psi = (d^c_{X\times B}\psi)_{\mathscr{H}} + \sum_{i=1}^{r}(d^c_{X\times B}\psi)(\xi_i^P - \xi_i^X)\theta_i = d^c_Y\psi - \langle d^c_X\psi, \theta\rangle. \tag{73}$$

We thus compute, on Z,

$$\begin{aligned}(d_Y d^c_Y\psi)_{(x,b)} &= d_Z\Bigl(d^c_X\psi + \sum_{j=1}^{r} d^c_X\psi(\xi_j^X)\theta_j + d^c_B\psi\Bigr) \\ &= d_Z d^c_X\psi + \sum_{j=1}^{r} d_Z(d^c_X\psi(\xi_j^X))\theta_j \\ &\qquad + \sum_{j=1}^{r} d^c_X\psi_b(\xi_j^X)\Bigl(\sum_{a=1}^{k}\xi^j(p_a)\pi_B^*\omega_a\Bigr) + d_Z d^c_B\psi \\ &= d_X d^c_X\psi_b + d_B d^c_B\psi_x + \sum_{j=1}^{r} d_Z(d^c_X\psi(\xi_j^X))\wedge\theta_j \\ &\qquad + \sum_{a=1}^{k} d^c_X\psi(p_a^X)\pi_B^*\omega_a + d_B d^c_X\psi + d_X d^c_B\psi,\end{aligned} \tag{74}$$

where for the third equality we used (22), as well as the identities $d_P d^c_X\psi = d_B d^c_X\psi$ and $d_P d^c_B\psi = d_B d^c_B\psi$ (which follow from the identification (25)). Using (27) and (44), we derive, from (72) and (74),

$$\Delta^Y_{\tilde\omega}(\psi)(x,b) = (\Delta^X_\omega\psi_b)(x) + (\Delta^B_{\omega_B(x)}\psi_x)(b) - \sum_{a=1}^{k}\frac{n_a}{\langle\mu_\omega, p_a\rangle + c_a}(d^c_X\psi_b)(p_a^X),$$

where, we recall, for a fixed $x \in X$ we have set $\omega_B(x) := \sum_{a=1}^{k}(\langle p_a, \mu_\omega\rangle + c_a)\omega_a$, and p_a^X denotes the vector field on X corresponding to $p_a \in \mathfrak{t}$. The first equality in the lemma follows from the identity (47), keeping in mind that, for any smooth function u on Δ and any $\mathbb{T}$–invariant smooth function ϕ on X,

$$g_\omega(d(u(\mu_\omega)), d\phi) = \sum_{i=1}^{r} u_{,i}(\mu_\omega)d^c\phi(\xi_i).$$

Now we establish the expression of the corresponding Lichnerowicz operators. Recall that (see eg [41])

$$\mathbb{L}^Y_{\tilde\omega}\psi := \tfrac{1}{2}(\Delta^Y_{\tilde\omega})^2(\psi) + \delta_{\tilde\omega}(\rho_{\tilde\omega}(d^c_Y\psi)). \tag{75}$$

Using the decomposition of $\Delta^Y_{\tilde{\omega}}$ we have just established,

$$(\Delta^Y_{\tilde{\omega}})^2(\psi) = (\Delta^X_{\omega,p})^2(\psi_b) + (\Delta^B_x)^2(\psi_x) + \Delta^X_{\omega,p}(\Delta^B_x(\psi_x)) + \Delta^B_x(\Delta^X_{\omega,p}(\psi_b)). \tag{76}$$

It remains to compute the Ricci term in (75). From (43),

$$\rho_{\tilde{\omega}} = \rho_{\omega,p} + \pi_B^*\rho_{\omega_B} + \frac{1}{2}\sum_{a=1}^k \Delta^X_{\omega,p}(\langle \mu_\omega, p_a^X\rangle)\pi_B^*\omega_a \\ + \sum_{j=1}^r d_X\big(d^c_X\big(\kappa - \tfrac{1}{2}\log p(\mu_\omega)\big)(\xi_j^X)\big) \wedge \theta_j, \tag{77}$$

where $\rho_{\omega,p} := \rho_\omega - \frac{1}{2} d_X d^c_X \log p(\mu_\omega)$ is the Ricci form of the weighted volume form $p(\mu_\omega)\omega^{[m]}$. Using integration by parts, for any $\mathbb{T}_Y$–invariant smooth test function ϕ on Y, seen as a $\mathbb{T}_X$ and $\mathbb{T}_P$–invariant function on $Z = X \times P$ via (25),

$$\begin{aligned} &\int_Z \phi\,\delta_{\tilde{\omega}}(\rho_{\tilde{\omega}}(d^c_Y\psi))\tilde{\omega}^{[n+m]} \wedge \theta^{\wedge r} \\ &= -\int_Z \rho_{\tilde{\omega}}(d_Y\phi, d^c_Y\psi)\tilde{\omega}^{[n+m]} \wedge \theta^{\wedge r} \\ &= \int_Z \rho_{\tilde{\omega}} \wedge d_Y\phi \wedge d^c_Y\psi \wedge \tilde{\omega}^{[n+m-2]} \wedge \theta^{\wedge r} \\ &\qquad - \frac{1}{2}\int_Z \mathrm{Scal}(\tilde{\omega})\tilde{g}_{\tilde{\omega}}(d_Y\phi, d_Y\psi)\tilde{\omega}^{[n+m]} \wedge \theta^{\wedge r} \\ &= \int_Z \rho_{\tilde{\omega}} \wedge d_Y\phi \wedge d^c_Y\psi \wedge \tilde{\omega}^{[n+m-2]} \wedge \theta^{\wedge r} \\ &\qquad - \frac{1}{2}\int_Z \left(\frac{\mathrm{Scal}_p(\omega)}{p(\mu_\omega)} + q(\mu_\omega)\right) d_Y\phi \wedge d^c_Y\psi \wedge \tilde{\omega}^{[n+m-1]} \wedge \theta^{\wedge r}. \end{aligned} \tag{78}$$

From the above formula, using (44), (73) and (77), we compute (after some straightforward but long algebraic manipulations and integration by parts over X and B)

$$\begin{aligned} &\delta^Y_{\tilde{\omega}}(\rho_{\tilde{\omega}}(d^c_Y\psi)) \\ &= \delta^X_{\omega,p}(\rho_{\omega,p}(d^c_X\psi)) + \delta^B_{\omega_B(x)}(\rho_{\omega_B}(d^c_B\psi)) \\ &\qquad + \frac{1}{2}\sum_{a=1}^k \frac{q(\mu_\omega)}{\langle \mu_\omega, p_a\rangle + c_a}\Delta^B_{\omega_a}(\psi) \\ &\qquad + \frac{1}{2}\sum_{a=1}^k \frac{(n_a - 1)}{(\langle \mu_\omega, p_a\rangle + c_a)^2}\Delta^X_{\omega,p}(\langle \mu_\omega, p_a\rangle)\Delta^B_{\omega_a}(\psi_x) \\ &\qquad + \sum_{a,b=1}^k \frac{n_b}{(\langle \mu_\omega, p_a\rangle + c_a)(\langle \mu_\omega, p_b\rangle + c_b)}\Delta^X_{\omega,p}(\langle \mu_\omega, p_b\rangle)\Delta^B_{\omega_a}(\psi_x). \end{aligned} \tag{79}$$

Combining (75), (76) and (79) yields the desired expression.

The expression for $(\mathbb{H}^{\tilde\rho}_{\tilde\omega,1})^Y(\psi)$ is obtained by similar arguments, using that

$$\begin{aligned}(\mathbb{H}^{\tilde\rho}_{\tilde\omega,1})^Y(\psi) &= \langle\tilde\rho, d_Y d_Y^c\psi\rangle_{\tilde\omega} + \langle d_Y \operatorname{tr}_{\tilde\omega}(\tilde\rho), d_Y\psi\rangle_{\tilde\omega} \\ &= -\operatorname{tr}_{\tilde\omega}(\tilde\rho)\Delta^Y_{\tilde\omega}(\psi) - \frac{\tilde\rho\wedge d_Y d_Y^c\psi\wedge\tilde\omega^{[n+m-2]}}{\tilde\omega^{[n+m]}} \\ &\qquad\qquad + \frac{d_Y \operatorname{tr}_{\tilde\omega}(\tilde\rho)\wedge d_Y^c\psi\wedge\tilde\omega^{[n+m-1]}}{\tilde\omega^{[n+m]}}. \qquad \square\end{aligned}$$

Appendix B Weighted Futaki invariants

On a smooth Fano manifold $(X,\mathbb{T})$ as in the setting and notation of Section 2, we further relate the weighted Futaki obstruction $\mathrm{Fut}_{v,w}=0$ (see (3)) with weights $v(\mu)$ and $w(\mu)$ as in Proposition 1 with the Futaki-type obstructions studied by Tian and Zhu [71] in the case of Kähler–Ricci solitons (ie when $v=e^{\langle\xi,\mu\rangle}$):

Lemma B.1 *Let $(X,\mathbb{T})$ be a smooth Fano manifold $(X,\mathbb{T})$ with canonically normalized momentum polytope Δ, and $v>0$ and w smooth functions on Δ as in Proposition 1. Then, for any $\mathbb{T}$–invariant Kähler metric $\omega\in 2\pi c_1(X)$ with momentum map μ_ω and $\mathbb{T}$–invariant Ricci potential h (ie $\rho_\omega-\omega=\frac{1}{2}dd^ch$), the weighted Futaki invariant $\mathrm{Fut}_{v,w}$ introduced in (3) satisfies*

$$\mathrm{Fut}_{v,w}(\ell_\zeta)=\int_X\bigl(\mathcal{L}_{J\zeta}(\log v(\mu_\omega)-h)\bigr)v(\mu_\omega)\omega^{[m]}=-2\int_X\langle\zeta,\mu_\omega\rangle v(\mu_\omega)\omega^{[m]}$$

for $\ell_\zeta=\langle\zeta,\mu\rangle+a$ and $\zeta\in\mathfrak{t}$.

Proof We have

$$\begin{aligned}\int_X\bigl(\mathcal{L}_{J\zeta}(\log v(\mu_\omega)-h)\bigr)v(\mu_\omega)\omega^{[m]} &= \int_X g_\omega\bigl(d\ell_\zeta, d\log(v(\mu_\omega)-h)\bigr)v(\mu_\omega)\omega^{[m]} \\ &= \int_X \ell_\zeta\bigl(\Delta_{\omega,v}(\log v(\mu_\omega)-h)\bigr)v(\mu_\omega)\omega^{[m]} \\ &= \int_X \ell_\zeta\bigl(\mathrm{Scal}_v(\omega)-w(\mu_\omega)\bigr)\omega^{[m]}=\mathrm{Fut}_{v,w}(\ell_\zeta),\end{aligned}$$

where for the last equality we have used (13). The second equality in the lemma follows from the first, the second relation in (11) and integration by parts. $\square$

References

[1] **V Apostolov**, *The Kähler geometry of toric manifolds* (2022) arXiv 2208.12493

[2] **V Apostolov**, **D M J Calderbank**, *The CR geometry of weighted extremal Kähler and Sasaki metrics*, Math. Ann. 379 (2021) 1047–1088 MR Zbl

[3] **V Apostolov**, **D M J Calderbank**, **P Gauduchon**, *Hamiltonian 2–forms in Kähler geometry, I: General theory*, J. Differential Geom. 73 (2006) 359–412 MR Zbl

[4] **V Apostolov**, **D M J Calderbank**, **P Gauduchon**, **C W Tønnesen-Friedman**, *Hamiltonian 2–forms in Kähler geometry, II: Global classification*, J. Differential Geom. 68 (2004) 277–345 MR Zbl

[5] **V Apostolov**, **D M J Calderbank**, **P Gauduchon**, **C W Tønnesen-Friedman**, *Hamiltonian 2–forms in Kähler geometry, IV: Weakly Bochner-flat Kähler manifolds*, Comm. Anal. Geom. 16 (2008) 91–126 MR Zbl

[6] **V Apostolov**, **D M J Calderbank**, **P Gauduchon**, **C W Tønnesen-Friedman**, *Extremal Kähler metrics on projective bundles over a curve*, Adv. Math. 227 (2011) 2385–2424 MR Zbl

[7] **V Apostolov**, **D M J Calderbank**, **E Legendre**, *Weighted K–stability of polarized varieties and extremality of Sasaki manifolds*, Adv. Math. 391 (2021) art. id. 107969 MR Zbl

[8] **C Arezzo**, **F Pacard**, **M Singer**, *Extremal metrics on blowups*, Duke Math. J. 157 (2011) 1–51 MR Zbl

[9] **M F Atiyah**, *Convexity and commuting Hamiltonians*, Bull. London Math. Soc. 14 (1982) 1–15 MR Zbl

[10] **R J Berman**, *K–polystability of $\mathbb{Q}$–Fano varieties admitting Kähler–Einstein metrics*, Invent. Math. 203 (2016) 973–1025 MR Zbl

[11] **R J Berman**, **B Berndtsson**, *Convexity of the K–energy on the space of Kähler metrics and uniqueness of extremal metrics*, J. Amer. Math. Soc. 30 (2017) 1165–1196 MR Zbl

[12] **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

[13] **R J Berman**, **T Darvas**, **C H Lu**, *Convexity of the extended K–energy and the large time behavior of the weak Calabi flow*, Geom. Topol. 21 (2017) 2945–2988 MR Zbl

[14] **R J Berman**, **T Darvas**, **C H Lu**, *Regularity of weak minimizers of the K–energy and applications to properness and K–stability*, Ann. Sci. Éc. Norm. Supér. 53 (2020) 267–289 MR Zbl

[15] **R J Berman**, **D W Nyström**, *Complex optimal transport and the pluripotential theory of Kähler–Ricci solitons*, preprint (2014) arXiv 1401.8264

[16] **Z Błocki**, *On geodesics in the space of Kähler metrics*, from "Advances in geometric analysis" (S Janeczko, J Li, D H Phong, editors), Adv. Lect. Math. 21, International, Somerville, MA (2012) 3–19 MR Zbl

[17] **Z Błocki**, **S Kołodziej**, *On regularization of plurisubharmonic functions on manifolds*, Proc. Amer. Math. Soc. 135 (2007) 2089–2093 MR Zbl

[18] **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

[19] **C P Boyer**, **K Galicki**, **S R Simanca**, *Canonical Sasakian metrics*, Comm. Math. Phys. 279 (2008) 705–733 MR Zbl

[20] **C P Boyer**, **C W Tønnesen-Friedman**, *The Sasaki join, Hamiltonian 2–forms, and constant scalar curvature*, J. Geom. Anal. 26 (2016) 1023–1060 MR Zbl

[21] **E Calabi**, *Extremal Kähler metrics*, from "Seminar on differential geometry" (S T Yau, editor), Ann. of Math. Stud. 102, Princeton Univ. Press (1982) 259–290 MR Zbl

[22] **X Chen**, *The space of Kähler metrics*, J. Differential Geom. 56 (2000) 189–234 MR Zbl

[23] **X Chen**, **J Cheng**, *On the constant scalar curvature Kähler metrics, II: Existence results*, J. Amer. Math. Soc. 34 (2021) 937–1009 MR Zbl

[24] **X Chen**, **M Paun**, **Y Zeng**, *On deformation of extremal metrics*, preprint (2015) arXiv 1506.01290

[25] **J Chu**, **V Tosatti**, **B Weinkove**, *On the $C^{1,1}$ regularity of geodesics in the space of Kähler metrics*, Ann. PDE 3 (2017) art. id. 15 MR Zbl

[26] **A S Dancer**, **M Y Wang**, *On Ricci solitons of cohomogeneity one*, Ann. Global Anal. Geom. 39 (2011) 259–292 MR Zbl

[27] **T Darvas**, *The Mabuchi geometry of finite energy classes*, Adv. Math. 285 (2015) 182–219 MR Zbl

[28] **T Darvas**, *The Mabuchi completion of the space of Kähler potentials*, Amer. J. Math. 139 (2017) 1275–1313 MR Zbl

[29] **T Darvas**, **Y A Rubinstein**, *Tian's properness conjectures and Finsler geometry of the space of Kähler metrics*, J. Amer. Math. Soc. 30 (2017) 347–387 MR Zbl

[30] **V Datar**, **G Székelyhidi**, *Kähler–Einstein metrics along the smooth continuity method*, Geom. Funct. Anal. 26 (2016) 975–1010 MR Zbl

[31] **T Delcroix**, *K–stability of Fano spherical varieties*, Ann. Sci. Éc. Norm. Supér. 53 (2020) 615–662 MR Zbl

[32] **F Delgove**, *Kähler–Ricci solitons on horospherical variety*, preprint (2017) arXiv 1710.05623

[33] **R Dervan**, *Relative K–stability for Kähler manifolds*, Math. Ann. 372 (2018) 859–889 MR Zbl

[34] **R Dervan**, **J Ross**, *K–stability for Kähler manifolds*, Math. Res. Lett. 24 (2017) 689–739 MR Zbl

[35] **R Dervan**, **L M Sektnan**, *Optimal symplectic connections on holomorphic submersions*, Comm. Pure Appl. Math. 74 (2021) 2132–2184 MR Zbl

[36] **S K Donaldson**, *Lower bounds on the Calabi functional*, J. Differential Geom. 70 (2005) 453–472 MR Zbl

[37] **S K Donaldson**, *Kähler geometry on toric manifolds, and some other manifolds with large symmetry*, from "Handbook of geometric analysis" (L Ji, P Li, R Schoen, L Simon, editors), Adv. Lect. Math. 7, International, Somerville, MA (2008) 29–75 MR Zbl

[38] **A Fujiki**, *On automorphism groups of compact Kähler manifolds*, Invent. Math. 44 (1978) 225–258 MR Zbl

[39] **A Futaki**, **H Ono**, *Conformally Einstein–Maxwell Kähler metrics and structure of the automorphism group*, Math. Z. 292 (2019) 571–589 MR Zbl

[40] **A Futaki**, **H Ono**, **G Wang**, *Transverse Kähler geometry of Sasaki manifolds and toric Sasaki–Einstein manifolds*, J. Differential Geom. 83 (2009) 585–635 MR Zbl

[41] **P Gauduchon**, *Calabi's extremal metrics: an elementary introduction*, lecture notes, in preparation

[42] **J P Gauntlett**, **D Martelli**, **J Sparks**, **D Waldram**, *A new infinite class of Sasaki–Einstein manifolds*, Adv. Theor. Math. Phys. 8 (2004) 987–1000 MR Zbl

[43] **V Guedj**, **A Zeriahi**, *The weighted Monge–Ampère energy of quasiplurisubharmonic functions*, J. Funct. Anal. 250 (2007) 442–482 MR Zbl

[44] **V Guillemin**, **S Sternberg**, *Convexity properties of the moment mapping*, Invent. Math. 67 (1982) 491–513 MR Zbl

[45] **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

[46] **Y Hashimoto**, *Existence of twisted constant scalar curvature Kähler metrics with a large twist*, Math. Z. 292 (2019) 791–803 MR Zbl

[47] **W He**, *On Calabi's extremal metric and properness*, Trans. Amer. Math. Soc. 372 (2019) 5595–5619 MR Zbl

[48] **W He**, **J Li**, *Geometric pluripotential theory on Sasaki manifolds*, J. Geom. Anal. 31 (2021) 1093–1179 MR Zbl

[49] **T Hisamoto**, *Stability and coercivity for toric polarizations*, preprint (2016) arXiv 1610.07998

[50] **E Inoue**, *Constant μ–scalar curvature Kähler metric–formulation and foundational results*, J. Geom. Anal. 32 (2022) art. id. 145 MR Zbl

[51] **D Jerison**, **J M Lee**, *Extremals for the Sobolev inequality on the Heisenberg group and the CR Yamabe problem*, J. Amer. Math. Soc. 1 (1988) 1–13 MR Zbl

[52] **S Jubert**, *A Yau–Tian–Donaldson correspondence on a class of toric fibrations*, preprint (2021) arXiv 2108.12297

[53] **N Koiso**, *On rotationally symmetric Hamilton's equation for Kähler–Einstein metrics*, from "Recent topics in differential and analytic geometry" (T Ochiai, editor), Adv. Stud. Pure Math. 18, Academic, Boston, MA (1990) 327–337 MR Zbl

[54] **A Lahdili**, *Automorphisms and deformations of conformally Kähler, Einstein–Maxwell metrics*, J. Geom. Anal. 29 (2019) 542–568 MR Zbl

[55] **A Lahdili**, *Kähler metrics with constant weighted scalar curvature and weighted K–stability*, Proc. Lond. Math. Soc. 119 (2019) 1065–1114 MR Zbl

[56] **A Lahdili**, *Convexity of the weighted Mabuchi functional and the uniqueness of weighted extremal metrics*, preprint (2020) arXiv 2007.01345

[57] **C Li**, *G–uniform stability and Kähler–Einstein metrics on Fano varieties*, Invent. Math. 227 (2022) 661–744 MR Zbl

[58] **T Mabuchi**, *Kähler–Einstein metrics for manifolds with nonvanishing Futaki character*, Tohoku Math. J. 53 (2001) 171–182 MR Zbl

[59] **T Mabuchi**, *Multiplier Hermitian structures on Kähler manifolds*, Nagoya Math. J. 170 (2003) 73–115 MR Zbl

[60] **T Mabuchi**, **Y Nakagawa**, *New examples of Sasaki–Einstein manifolds*, Tohoku Math. J. 65 (2013) 243–252 MR Zbl

[61] **D Martelli**, **J Sparks**, **S-T Yau**, *Sasaki–Einstein manifolds and volume minimisation*, Comm. Math. Phys. 280 (2008) 611–673 MR Zbl

[62] **Y Odaka**, *A generalization of the Ross–Thomas slope theory*, Osaka J. Math. 50 (2013) 171–185 MR Zbl

[63] **F Podestà**, **A Spiro**, *Kähler–Ricci solitons on homogeneous toric bundles*, J. Reine Angew. Math. 642 (2010) 109–127 MR Zbl

[64] **S Semmes**, *Complex Monge–Ampère and symplectic manifolds*, Amer. J. Math. 114 (1992) 495–550 MR Zbl

[65] **Z Sjöström Dyrefelt**, *K–semistability of cscK manifolds with transcendental cohomology class*, J. Geom. Anal. 28 (2018) 2927–2960 MR Zbl

[66] **Z Sjöström Dyrefelt**, *On K–polystability of cscK manifolds with transcendental cohomology class*, Int. Math. Res. Not. 2020 (2020) 2769–2817 MR Zbl

[67] **J Stoppa**, *K–stability of constant scalar curvature Kähler manifolds*, Adv. Math. 221 (2009) 1397–1408 MR Zbl

[68] **J Stoppa**, **G Székelyhidi**, *Relative K–stability of extremal metrics*, J. Eur. Math. Soc. 13 (2011) 899–909 MR Zbl

[69] **A Swann**, *Twisting Hermitian and hypercomplex geometries*, Duke Math. J. 155 (2010) 403–431 MR Zbl

[70] **G Tian**, *Kähler–Einstein metrics with positive scalar curvature*, Invent. Math. 130 (1997) 1–37 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] **X Wang**, *Height and GIT weight*, Math. Res. Lett. 19 (2012) 909–926 MR Zbl

[73] **X-J Wang**, **X Zhu**, *Kähler–Ricci solitons on toric manifolds with positive first Chern class*, Adv. Math. 188 (2004) 87–103 MR Zbl

Département de Mathématiques, Université du Québec à Montréal
Montréal, QC, Canada

Current address: *Institute of Mathematics and Informatics, Bulgarian Academy of Sciences*
Sofia, Bulgaria

Département de Mathématiques, Université du Québec à Montréal
Montréal, QC, Canada

Current address: *Institut de Mathématiques de Toulouse*
Toulouse, France

Department of Mathematics, Aarhus University
Aarhus, Denmark

apostolov.vestislav@uqam.ca, simonjubert@gmail.com,
lahdili.abdellah@gmail.com

Proposed: Gang Tian
Seconded: Frances Kirwan, Simon Donaldson

Received: 17 April 2021
Revised: 4 December 2021

msp

Geometry & Topology **27**:8 (2023) 3303–3360
DOI: 10.2140/gt.2023.27.3303
Published: 9 November 2023

Anosov representations with Lipschitz limit set

MARIA BEATRICE POZZETTI
ANDRÉS SAMBARINO
ANNA WIENHARD

We study Anosov representations whose limit set has intermediate regularity, namely is a Lipschitz submanifold of a flag manifold. We introduce an explicit linear functional, the unstable Jacobian, whose orbit growth rate is integral on this class of representations. We prove that many interesting higher-rank representations, including Θ–positive representations, belong to this class, and establish several applications to rigidity results on the orbit growth rate in the symmetric space.

22E40, 51F30

1 Introduction

Let $\Gamma \subset \mathsf{PGL}_d(\mathbb{R})$ be a discrete subgroup. Following Guivarc'h, Benoist [4] has shown that if Γ contains a proximal element and acts irreducibly on $\mathbb{R}^d$ then its action on projective space $\mathbb{P}(\mathbb{R}^d)$ has a smallest closed invariant set. This is usually called *Benoist's limit set* or simply *the limit set* of Γ on $\mathbb{P}(\mathbb{R}^d)$ and denoted by L_Γ.

In contrast with the negatively curved situation, the limit set of a subgroup Γ whose Zariski closure has rank ≥ 2 need not be a fractal object. Examples of infinite covolume Zariski-dense groups whose limit set is a proper C^1–submanifold arise in the study of strictly convex divisible sets (see Benoist [5]) and of Hitchin representations (see Labourie [35]). Lately, more examples of subgroups with this property were found by Pozzetti, Sambarino and Wienhard [39] and Zhang and Zimmer [48].

Intermediate phenomena also occur. For example, the limit set of the direct sum $(\rho,\eta)\colon \pi_1 S \to \mathsf{PSL}_2(\mathbb{R})\times\mathsf{PSL}_2(\mathbb{R})$ of the holonomies of two hyperbolizations of a closed topological surface S is a Lipschitz circle that is never a C^1–submanifold of the product $\mathbb{S}^1\times\mathbb{S}^1=\mathsf{G}/\mathsf{B}$ unless the two hyperbolizations are conjugated. Lipschitz limit sets more generally occur for maximal representations (see Burger, Iozzi and Wienhard [12]), Quasi-Fuchsian AdS representations (see Barbot and Mérigot [3]), and $\mathbb{H}^{p,q}$–convex–cocompact representations; see Danciger, Guéritaud and Kassel [16].

We provide the first systematic investigation of this intermediate phenomenon — its main object are discrete groups whose limit set is a Lipschitz manifold. We will restrict our investigation to the class of Anosov subgroups, a robust and rich class of strongly undistorted subgroups of semisimple Lie groups; see Section 2.2 for the precise definition.

For discrete subgroups Γ of $\mathsf{SO}(1,n)$, Sullivan [46] established a beautiful relation between a geometric invariant of the limit set L_Γ, its Hausdorff dimension, and a dynamical invariant for the action of Γ on the symmetric space $\mathbb{H}^n$, the orbit growth rate. This was further used by Bowen [7] to prove a strong rigidity result: for fundamental groups of surfaces acting on $\mathbb{H}^3$, the Hausdorff dimension of the limit set is minimal if and only if the limit set is C^1 and Γ preserves a totally geodesic copy of $\mathbb{H}^2$ on which it acts cocompactly. When G has higher rank, the situation is more complicated as one can additionally consider orbit growth rates with respect to different linear functionals φ (as in, for example, Quint [41]). It is a challenging problem to understand which functionals φ have orbit growth rate that carries geometric information on the group Γ or on its limit set L_Γ.

Our main contribution is to single out an explicit linear functional, the *unstable Jacobian*, whose critical exponent is integral on Anosov subgroups whose limit set is a Lipschitz submanifold. In order to prove this we import ideas from nonconformal dynamics, such as the study of the affinity exponent, to the setting of Anosov groups, and use the Anosov property, together with ideas from geometric group theory, to establish a strengthening of the theory of Patterson–Sullivan densities developed by Quint; these two results are of independent interest. We then showcase the strength of our main result by applying it to several well-studied classes of representations: maximal representations, $\mathbb{H}^{p,q}$–convex–cocompact subgroups and Θ–positive representations.

The unstable Jacobian and the affinity exponent

We now introduce some notation useful to explain more precisely our results. We denote by

$$\mathsf{E} = \left\{ \underline{a} = (a_1, \dots, a_d) \in \mathbb{R}^d \;\middle|\; \sum_i a_i = 0 \right\}$$

the Cartan subspace of the Lie group $\mathsf{PGL}_d(\mathbb{R})$, by

$$\mathsf{a}_i(\underline{a}) = a_i - a_{i+1}$$

the i^{th} simple root and by $\mathsf{E}^+ \subset \mathsf{E}$ the Weyl chamber whose associated set of simple roots is $\Pi = \{\mathsf{a}_i : i \in [\![1, d-1]\!]\}$. Let $a : \mathsf{PGL}_d(\mathbb{R}) \to \mathsf{E}^+$ be the *Cartan projection* with respect to the choice of a scalar product τ. Concretely, $a(g) = (\log \sigma_1(g), \dots, \log \sigma_d(g))$, where the $\sigma_i(g)$ denote the *singular values* of the matrix g, the square roots of the eigenvalues of the matrix gg^*, where g^* is the adjoint operator of g with respect to τ.

Given a discrete subgroup $\Gamma < \mathsf{PGL}_d(\mathbb{R})$, the *critical exponent* of a linear form $\varphi \in \mathsf{E}^*$, denoted by $h_\Gamma(\varphi)$, is defined as

$$h_\Gamma(\varphi) := \lim_{T \to \infty} \frac{\log \#\{\gamma \in \Gamma \mid \varphi(a(\gamma)) < T\}}{T}.$$

We introduce the p^{th} *unstable Jacobian* $\mathcal{J}_p^u \in \mathsf{E}^*$, defined by

$$\mathcal{J}_p^u = (p+1)\omega_{\mathsf{a}_1} - \omega_{\mathsf{a}_{p+1}},$$

where $\omega_{\mathsf{a}_p}(\underline{a}) = \sum_1^p a_i$ is the fundamental weight relative to the p^{th} simple root a_p. Our main result is:

Theorem A *Let $\Gamma < \mathsf{PSL}_d(\mathbb{R})$ be a strongly irreducible, projective Anosov subgroup whose limit set $\mathsf{L}_\Gamma < \mathbb{P}(\mathbb{R}^d)$ is a Lipschitz submanifold of dimension p. Then*

$$h_\Gamma(\mathcal{J}_p^u) = 1.$$

If $p = 1$ the same holds, replacing strong irreducibility with weak irreducibility.[1]

[1] We say that a subgroup $\Gamma < \mathsf{PSL}_d(\mathbb{R})$ is *weakly irreducible* if the vector space $\mathrm{span}(\mathsf{L}_\Gamma)$ is $\mathbb{R}^d$.

A similar result was proven, in the context of fundamental groups of compact strictly convex projective manifolds, by Potrie and Sambarino [38, Theorem B]; our approach is entirely different and, since we require less regularity, its scope of application is considerably broader. Note that, up to postcomposing with a suitable linear representation, any Anosov representation can be turned into a projective Anosov representation.

We prove the two inequalities in Theorem A as corollaries of two different results that are applicable in more general settings. We focus first on the lower bound on the critical exponent (Corollary 1.1) that follows from a general result on the Hausdorff dimension of limit sets (of projective Anosov representations).

An important step in the proof is the study, in the context of Anosov representations, of the *affinity exponent*, a key notion from nonconformal dynamics that first appeared in Kaplan and Yorke [31] and Douady and Oesterlé [19], and played a prominent role in Falconer's work [21]. More specifically, for a discrete subgroup $\Gamma < \mathsf{PSL}_d(\mathbb{R})$, we consider the *piecewise* Dirichlet series defined, for $p \in \mathbb{N}$ and $s \in [p-1, p]$, by

$$\Phi_\Gamma^{\mathsf{Aff}}(s) = \sum_{\gamma \in \Gamma} \left(\frac{\sigma_2}{\sigma_1}(\gamma) \cdots \frac{\sigma_p}{\sigma_1}(\gamma) \right) \left(\frac{\sigma_{p+1}}{\sigma_1}(\gamma) \right)^{s-(p-1)}.$$

The affinity exponent is the critical exponent of this series:

$$h_\Gamma^{\mathsf{Aff}} := \inf\{s : \Phi_\Gamma^{\mathsf{Aff}}(s) < \infty\} = \sup\{s : \Phi_\Gamma^{\mathsf{Aff}}(s) = \infty\} \in (0, \infty].$$

Our second main result is (see Section 3 for a statement for arbitrary local fields):

Theorem B *Let $\Gamma < \mathsf{PGL}_d(\mathbb{R})$ be projective Anosov, then*

$$\dim_{\mathrm{Hff}}(\mathsf{L}_\Gamma) \le h_\Gamma^{\mathsf{Aff}}.$$

It is easy to deduce from Theorem B relations between the Hausdorff dimension of the limit set of a projective Anosov subgroup and the orbit growth rate with respect to explicit linear functionals on the Weyl chamber. Since the quantity $h_\Gamma(\mathcal{J}_p^u)$ appearing in Theorem A is also the critical exponent of the Dirichlet series

$$s \mapsto \sum_{\gamma \in \Gamma} \left(\frac{\sigma_1 \cdots \sigma_{p+1}}{\sigma_1^{p+1}}(\gamma) \right)^s,$$

we get:

Corollary 1.1 *Let $\Gamma < \mathsf{PGL}_d(\mathbb{R})$ be projective Anosov and assume furthermore that $\dim_{\mathrm{Hff}}(\mathsf{L}_\Gamma)) \ge p$. Then*

$$\dim_{\mathrm{Hff}}(\mathsf{L}_\Gamma) \le p h_\Gamma(\mathcal{J}_p^u).$$

Observe that $\mathcal{J}_1^u = \mathsf{a}_1$ and thus, whenever $\dim_{\mathrm{Hff}}(\mathsf{L}_\Gamma) \geq 1$, we obtain as a consequence the results of Glorieux, Monclair and Tholozan [25, Theorem 4.1] and Pozzetti, Sambarino and Wienhard [39, Proposition 4.1].

Existence of Patterson–Sullivan measures

The second inequality in Theorem A follows from an improvement on a result by Quint [42, théorème 8.1] concerning the relation between critical exponents and the existence of (Γ, φ)–Patterson–Sullivan measures.

Given a set $\Theta \subset \Pi$ of simple roots, we denote by $\mathcal{F}_\Theta$ the associated partial flag manifold, which consists of the space of flags of subspaces of dimension indexed by Θ. We denote by E_Θ the Levi subspace of E defined by

$$\mathsf{E}_\Theta = \bigcap_{p \notin \Theta} \ker \mathsf{a}_p.$$

The restrictions of the fundamental weights $\{\omega_{\mathsf{a}_p}|_{\mathsf{E}_\Theta} : p \in \Theta\}$ span its dual $(\mathsf{E}_\Theta)^*$. Using the Iwasawa decomposition of $\mathsf{PGL}_d(\mathbb{R})$, Quint introduced an *Iwasawa cocycle*

$$\mathrm{b}_\Theta \colon \mathsf{PGL}_d(\mathbb{R}) \times \mathcal{F}_\Theta \to \mathsf{E}_\Theta$$

that is the higher-rank analog of the more-studied *Busemann cocycle* in negative curvature; see Quint [42, lemme 6.6] and Section 5.3 for the precise definition. With this notation at hand we can recall the definition of a (Γ, φ)–Patterson–Sullivan measure from [42]:

Definition 1.2 Given a discrete subgroup $\Gamma < \mathsf{PGL}_d(\mathbb{R})$ and $\varphi \in (\mathsf{E}_\Theta)^*$, a (Γ, φ)–*Patterson–Sullivan measure* on $\mathcal{F}_\Theta$ is a finite Radon measure μ such that, for every $g \in \Gamma$,

$$\frac{\mathrm{d}g_*\mu}{\mathrm{d}\mu}(x) = e^{-\varphi(\mathrm{b}_\Theta(g^{-1}, x))}.$$

Inspired by a classical result by Sullivan [46], Quint shows [42, théorème 8.1] that the existence of a (Γ, φ)–Patterson–Sullivan measure on $\mathcal{F}_\Theta$ gives an upper bound on a related critical exponent

$$h_\Gamma(\varphi + \rho_{\theta^c}) \leq 1. \tag{1-1}$$

Here ρ_{θ^c} is an explicit linear functional which is positive on the interior of the Weyl chamber and accounts for the possible growth along the fibers of the projection $\mathcal{F}_\Delta \to \mathcal{F}_\Theta$ [42, lemme 8.3]. In general, $h_\Gamma(\varphi + \rho_{\theta^c}) < h_\Gamma(\varphi)$, and thus Quint's result is not sharp enough for our purposes.

Using ideas from geometric group theory we show that, provided the group $\mathsf{\Gamma}$ is Anosov with respect to one of the roots in Θ, there is no contribution from the fibers.

Given $\Theta \subset \Pi$, define $\mathrm{i}\Theta = \{d-p : p \in \Theta\}$. Two points $(x, y) \in \mathcal{F}_\Theta \times \mathcal{F}_{\mathrm{i}\Theta}$ are *transverse* if, for every $p \in \Theta$, one has that $x^p \cap y^{d-p} = \{0\}$. A *complementary subspace of* $\mathcal{F}_\Theta$ is a subset of $\mathcal{F}_\Theta$ of the form

$$\{x \in \mathcal{F}_\Theta : x \text{ is not transverse to } y_0\}$$

for a given $y_0 \in \mathcal{F}_{\mathrm{i}\Theta}$. If $\Theta' \subset \Theta$ then we let $\pi_{\Theta,\Theta'} \colon \mathcal{F}_\Theta \to \mathcal{F}_{\Theta'}$ be the canonical projection.

Theorem C *Let $\mathsf{\Gamma} < \mathsf{PGL}_d(\mathbb{R})$ be projective Anosov and consider $\Theta \subset \Pi$ such that $\mathsf{a}_1 \in \Theta$. Let $\varphi \in (\mathsf{E}_\Theta)^*$. If there exists a $(\mathsf{\Gamma}, \varphi)$–Patterson–Sullivan measure on $\mathcal{F}_\Theta$ with support on $\pi^{-1}_{\Theta,\mathsf{a}_1}(\mathsf{L}_\mathsf{\Gamma})$ and not contained on a complementary subspace, then*

$$h_\mathsf{\Gamma}(\varphi) \leq 1.$$

We refer the reader to Section 5 and Theorem 5.14 for a version of Theorem C where the target group is an arbitrary semisimple group over a local field.

We provide the link between Theorems C and A in Section 6, where we establish that, if $\mathsf{\Gamma} < \mathsf{PSL}_d(\mathbb{R})$ is a projective Anosov subgroup whose limit set $\mathsf{L}_\mathsf{\Gamma}$ is a Lipschitz submanifold of dimension p, then there exists a $(\mathsf{\Gamma}, \mathcal{J}^u_p)$–Patterson–Sullivan measure on $\mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_p\}}$. In fact we explicitly construct such a measure using Rademacher's theorem and an explicit volume form on the almost everywhere defined tangent space to $\mathsf{L}_\mathsf{\Gamma}$ (Proposition 6.4).

Example 1.3 If $\rho \colon \pi_1 S \to \mathsf{PSp}(4,\mathbb{R})$ is a maximal representation (see Section 9 for the definition), the combination of Theorems B and C gives $h_{\rho(\pi_1 S)}(\mathsf{a}_2) = 1$, while Quint's result (1-1) becomes $h_{\rho(\pi_1 S)}(\omega_{\mathsf{a}_2}) \leq 1$. This latter inequality is implied by the former equality, and often far from being sharp: one can find representations ρ for which $h_{\rho(\pi_1 S)}(\omega_{\mathsf{a}_2})$ is arbitrarily small.

Theorem C is complementary to (and independent from) the Patterson–Sullivan theory for Anosov representations developed by Dey and Kapovich [18]. They only consider Patterson–Sullivan densities with respect to functionals φ that, as opposed to the unstable Jacobian, belong to $(\mathsf{E}_\theta)^*$, where the representation is assumed to be Anosov with respect to *all* elements of θ, and induce Finsler distances on the symmetric space; see also Ledrappier [36] for a different approach yielding similar results. A drawback of their approach is that they can only relate the critical exponent with a premetric induced from a Finsler distance on the symmetric space that is hard to compute. In contrast, we begin with a natural measure, supported on the limit set, which belongs to the Lebesgue measure class, find a suitable functional, the unstable Jacobian, turning the measure into a Patterson–Sullivan measure, and deduce from this geometric properties of the action of Γ on the symmetric space.

Intermediate regularity and C^1–dichotomy

The class of Anosov subgroups with Lipschitz limit sets is very rich, and includes the images of many well-studied classes of representations, such as maximal representations (see Burger, Iozzi and Wienhard [12] and Section 9), quasi-Fuchsian AdS representations (see Barbot and Mérigot [3]) and $\mathbb{H}^{p,q}$–convex–cocompact representations (see Danciger, Guéritaud and Kassel [16] and Section 8).

As another contribution of independent interest, we show that Θ–positive representations of fundamental groups of surfaces in $\mathsf{SO}(p,q)$ (see Guichard and Wienhard [27]) yield subgroups with this property. We refer the reader to Section 10 for the precise definition of Θ–positive representations. We will only[2] consider here the Θ–positive representations that are furthermore Θ–Anosov for $\Theta = \{\mathsf{a}_1, \dots, \mathsf{a}_{p-1}\}$. As a result, for each $k \in \Theta$, they admit a boundary map $\xi^k : \partial\Gamma \to \mathsf{Is}_k(\mathbb{R}^{p,q})$ parametrizing the limit set in the Grassmannian of k–dimensional isotropic subspaces. In Section 10 we prove:

Theorem D *Let $\rho\colon \Gamma \to \mathsf{SO}(p,q)$ be a Θ–Anosov representation that is Θ–positive. Then the images of the boundary maps $\xi^k : \partial\Gamma \to \mathsf{Is}_k(\mathbb{R}^{p,q})$ are C^1–submanifolds for each $1 \le k < p-1$; moreover $\xi^{p-1}(\partial\Gamma)$ is Lipschitz.*

We will prove the parts of Theorem D separately, in Corollary 10.4 and Proposition 10.5, respectively.

At least for representations of fundamental groups of surfaces, the regularity of the limit set on a given (maximal) flag space seems to be related to the position of the associated root among the Anosov roots. By definition, a simple root is an *Anosov root* (for a subgroup Γ) if its kernel intersects trivially the limit cone $\mathcal{L}_\Gamma$ of Γ. Among such roots one can consider the internal (every neighboring root in the Dynkin diagram is also an Anosov root) or boundary (connected to a root that nontrivially intersects $\mathcal{L}_\Gamma$) roots. For example, for a Θ–positive representation in $\mathsf{SO}(p,q)$, the roots $\{\mathsf{a}_1, \dots, \mathsf{a}_{p-2}\}$ are internal while a_{p-1} is the only boundary root.

[2] Guichard, Labourie and Wienhard announced that all Θ–positive representations are Θ–Anosov, so this should not pose any restriction.

The intermediate regularity (Lipschitz but not C^1) of limit sets for surface groups seems only to occur for boundary roots. For internal roots, we can prove a C^1–dichotomy, ruling out intermediate regularity in several interesting cases. More specifically we consider fundamental groups Γ of compact surfaces and study small deformations of representations of the form

$$\Gamma \to \mathsf{PSL}_2(\mathbb{R}) \xrightarrow{R} \mathsf{PSL}_d(\mathbb{R})$$

that are $\{\mathsf{a}_1, \mathsf{a}_2\}$–Anosov (this latter assumption can be rephrased as a proximality assumption on the linear representation R). For any such representation we have an explicit dichotomy: the associated limit set is either C^1 or not even Lipschitz (Corollary 7.8). We refer the reader to Section 7 for the precise statement of the dichotomy.

Entropy rigidity results

We conclude the introduction by discussing three well-studied classes of representations to which Theorem A applies. Interestingly, in all these cases, the mere information on the critical exponent of the unstable Jacobian provided by Theorem A allows us to obtain a sharp upper bound on the critical exponent for the action on the symmetric space endowed with the Riemannian distance function. In the case of Θ–positive representations this is even sufficient to prove that the bound is rigid: it is attained only on the specific Fuchsian locus, the generalization, in our setting, of Bowen's aforementioned result.

Maximal representations

Maximal representations are well-studied representations of fundamental groups of surfaces in Hermitian Lie groups $\mathsf{G}_{\mathbb{R}}$ that were introduced by Burger, Iozzi and Wienhard [12] through a cohomological invariant, the Toledo invariant. For these representations Theorem A applies, and gives:

Theorem 1.4 *Let $\mathsf{G}_{\mathbb{R}}$ be a classical simple Hermitian Lie group of tube type. Let $\rho\colon \Gamma \to \mathsf{G}_{\mathbb{R}}$ be a maximal representation, and let $\check{\mathsf{a}}$ denote the root associated to the stabilizer of a point in the Shilov boundary of $\mathsf{G}_{\mathbb{R}}$. Then $h_\rho(\check{\mathsf{a}}) = 1$.*

Concretely, in the case $\mathsf{G}_{\mathbb{R}} \in \{\mathsf{Sp}(2p, \mathbb{R}), \mathsf{SU}(p, p), \mathsf{SO}^*(4p)\}$, the root $\check{\mathsf{a}}$ computes the logarithm of the square of the middle eigenvalue, while for $\mathsf{G} = \mathsf{SO}_0(2, p)$ the root $\check{\mathsf{a}}$ is the first root, computing the logarithm of the first eigenvalue gap.

Theorem 1.4 also holds for the exceptional Hermitian Lie group of tube type if the representation is Zariski-dense, and we expect it to hold unconditionally. We refer the reader to Section 9 for a slightly more general statement, further explanations and consequences, in particular concerning a sharp upper bound on the exponential orbit growth rate for the action on the symmetric space (see Proposition 9.9).

$\mathbb{H}^{p,q}$–convex–cocompact representations

Generalizing work of Mess [37] and Barbot and Mérigot [3], Danciger, Guéritaud and Kassel [16] introduced a class of representations called $\mathbb{H}^{p,q}$*–convex–cocompact*. Here $\mathbb{H}^{p,q}$ is the *pseudo-Riemannian hyperbolic space*, consisting of negative lines in $\mathbb{P}(\mathbb{R}^d)$ for a fixed nondegenerate form Q of signature $(p, q+1)$. It follows then from [16, Theorem 1.11] that a projective Anosov subgroup $\Gamma < \mathsf{PO}(Q) = \mathsf{PO}(p, q+1)$ is $\mathbb{H}^{p,q}$*–convex–cocompact* if, for every pairwise distinct triple of points $x, y, z \in \mathsf{L}_\Gamma$, the restriction $Q|_{\langle x,y,z\rangle}$ has signature $(2, 1)$.

Consider a representation $\Lambda : \mathsf{PO}(p, 1) \to \mathsf{PO}(p, q+1)$ whose image stabilizes a $(p+1)$–dimensional subspace V of $\mathbb{R}^d$, where $Q|_V$ has signature $(p, 1)$. Endow the symmetric space $X_{p,q+1}$ with the $\mathsf{PO}(p, q+1)$–invariant Riemannian metric normalized so that the totally geodesic copy of $\mathbb{H}^p$ in $X_{p,q+1}$ stabilized by Λ has constant curvature -1.

Definition 1.5 For a subgroup $\Gamma < \mathsf{SO}(p, q+1)$ and $x_0 \in X_{p,q+1}$, denote by $h_\rho^{X_{p,q+1}}$ the critical exponent of the Dirichlet series

$$s \mapsto \sum_{\gamma\in\Gamma} e^{-sd(x_0,\rho(\gamma)x_0)}.$$

We have the upper bound:

Proposition 1.6 *Assume that $\partial\Gamma$ is homeomorphic to a $(p-1)$–dimensional sphere, and let $\Gamma < \mathsf{PO}(p, q+1)$ be strongly irreducible and $\mathbb{H}^{p,q}$–convex–cocompact. Then*

$$h_\rho^{X_{p,q+1}} \le p - 1.$$

We expect this upper bound to be rigid, namely the upper bound should only be attained at an inclusion of a cocompact lattice in $\mathsf{PO}(p, 1)$ preserving a totally geodesic copy of $\mathbb{H}^p$ of the type induced by Λ. However, only the case $p = 2$ is known; see Collier, Tholozan and Toulisse [15].

Section 8 contains more information on $\mathbb{H}^{p,q}$–convex–cocompact representations. In particular the relation with recent work by Glorieux and Monclair [24].

Θ–positive representations

Thanks to Theorem D, Theorem A also applies to Θ–positive representations of fundamental groups of surfaces in $\mathsf{SO}(p,q)$ and gives:

Corollary 1.7 *Let $\rho\colon \Gamma \to \mathsf{SO}(p,q)$ be a Θ–Anosov representation that is Θ–positive and weakly irreducible. Then $h_\rho(\mathsf{a}_k) = 1$ for every $k \le p-1$.*

Inspired by Potrie and Sambarino [38], we deduce from Corollary 1.7 a rigid upper bound for the critical exponent of the action of a positive representation on the Riemannian symmetric space $X_{p,q}$ (see Theorem 10.7). More precisely, we now normalize the $\mathsf{SO}(p,q)$–invariant Riemannian metric on $X_{p,q}$ so that the totally geodesic copy of $\mathbb{H}^2$ induced by the representation $\Lambda\colon \mathsf{SL}_2(\mathbb{R}) \to \mathsf{SO}(p,q)$ that stabilizes a subspace of $\mathbb{R}^d$ of signature $(p, p-1)$ has constant curvature -1. We consider the critical exponent in Definition 1.5 with this normalization of distance.

Theorem 1.8 *Let Γ be the fundamental group of a surface and let $\rho\colon \Gamma \to \mathsf{SO}(p,q)$ be Θ–positive. Then the critical exponent with respect to the Riemannian metric satisfies*

$$h_\rho^{X_{p,q}} \le 1.$$

Furthermore, if equality is achieved at a totally reducible representation η, then η splits as $W \oplus V$, W has signature $(p, p-1)$, $\eta|_W$ has Zariski closure the irreducible $\mathsf{PO}(2,1)$ in $\mathsf{PO}(p, p-1)$, and $\eta|_V$ lies in a compact group.

New arguments are needed with respect to [38], since the Anosov–Levi space of a Θ–positive representation has codimension one (instead of 0, which is the case treated in [38]); see Section 10.

Plan of the paper

In Section 2 we introduce some required preliminaries, and recall some needed results from Bochi, Potrie and Sambarino [6] and Pozzetti, Sambarino and Wienhard [39]. Section 3 deals with the affinity exponent and Hausdorff dimension for Anosov representations, and in it we prove Theorem B for any local field. Section 4 is a reminder of (more or less) standard definitions on semisimple algebraic groups over a local field. In Section 5 we recall objects from higher-rank Patterson–Sullivan theory and in Section 5.3 we prove Theorem 5.14 (a broader version of Theorem C). Section 6 completes the proof of Theorem A. The remaining sections deal with the applications of this result discussed in the introduction.

Acknowledgments

We would like to thank J-F Quint for pointing us to Falconer's work and suggesting to consider the affinity exponent.

Sambarino was partially financed by ANR DynGeo ANR-16-CE40-0025. Pozzetti and Wienhard acknowledge funding by the Deutsche Forschungsgemeinschaft within the priority program SPP 2026 *Geometry at infinity*, and from the National Science Foundation under grant 1440140 while they were in residence at the Mathematical Sciences Research Institute in Berkeley, California, during the semester of fall 2019. Wienhard acknowledges funding by the European Research Council under ERC Consolidator grant 614733, and by the Klaus Tschira Foundation. This work is supported by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy EXC-2181/1 – 390900948 (the Heidelberg STRUCTURES Cluster of Excellence).

2 Preliminaries

We recall in this section the notions we will need concerning Anosov representations and cone types. We refer the reader to [6; 39] for more details.

Throughout the paper $\mathbb{K}$ will denote a local field with absolute value $\|\cdot\|\colon \mathbb{K}\to\mathbb{R}^+$. If $\mathbb{K}$ is non-Archimedean, we require that $|\omega| = 1/q$ where ω denotes the *uniformizing element*, namely a generator of the maximal ideal of the valuation ring $\mathcal{O}$, and q is the cardinality of the residue field $\mathcal{O}/\omega\mathcal{O}$ (this is finite because $\mathbb{K}$ is, by assumption, local). This guarantees that the Hausdorff dimension of $\mathbb{P}^1(\mathbb{K})$ equals 1.

2.1 Singular values and Anosov representations into $\mathsf{PGL}_d(V_{\mathbb{K}})$

A $\mathbb{K}$–norm $\|\cdot\|$ on a $\mathbb{K}$ vector space $V_{\mathbb{K}}$ induces a norm on every exterior power of V; the angle between two vectors $\angle(v,w)$ is the unique number in $[0,\pi]$ such that

$$\sin\angle(v,w) := \frac{\|v\wedge w\|}{\|v\|\,\|w\|}.$$

Given two points $[v],[w]\in\mathbb{P}(V)$, we define their distance as

$$d([v],[w]) := \sin\angle(v,w),$$

and, given any two subspaces $P, Q < V$, we define their minimal angle as

$$\angle(P,Q) = \min_{v\in P\setminus\{0\}}\ \min_{w\in Q\setminus\{0\}} \angle(v,w).$$

An element $a \in \mathsf{GL}(V_{\mathbb{K}})$ is a *semihomothecy* (for a norm $\|\cdot\|$) if there exists an a–invariant $\mathbb{K}$–orthogonal[3] decomposition $V = V_1 \oplus \cdots \oplus V_k$ and $\sigma_1, \dots, \sigma_k \in \mathbb{R}_+$ such that, for every $i \in [\![1,k]\!]$ and every $v_i \in V_i$,

$$\|av_i\| = \sigma_i \|v_i\|.$$

The numbers σ_i are called the ratios of the semihomothecy a.

Following Quint [40, théorème 6.1], we fix a maximal abelian subgroup of diagonalizable matrices $A \subset \mathsf{GL}(V_{\mathbb{K}})$, a compact subgroup $K \subset \mathsf{GL}(V_{\mathbb{K}})$ such that if N is the normalizer of A in $\mathsf{GL}(V_{\mathbb{K}})$ then $N = (N \cap K)A$, and a $\mathbb{K}$–norm $\|\cdot\|$ on V preserved by K and such that A acts on V by semihomothecies. Let $e_1 \oplus \cdots \oplus e_d$ be the eigenlines of A (here $d = \dim V$) and choose the Weyl chamber A^+ consisting of those elements $a \in A$ whose corresponding semihomothecy ratios verify $\sigma_1(a) \geq \cdots \geq \sigma_d(a)$.

For every $g \in \mathsf{GL}(V_{\mathbb{K}})$ we choose a Cartan decomposition $g = k_g a_g l_g$ with a_g in A^+ and $k_g, l_g \in K$, and denote by

$$\sigma_1(g) \geq \sigma_2(g) \geq \cdots \geq \sigma_d(g)$$

the semihomothecy ratios of the Cartan projection $a_g \in A^+$ (these do not depend on the choice of the Cartan decomposition once K and $\|\cdot\|$ are fixed). In order to simplify notation we will often write $(\sigma_i/\sigma_j)(g) = \sigma_i(g)/\sigma_j(g)$.

We define, for $p \in [\![1, d-1]\!]$,

$$u_p(g) = k_g \cdot e_p \in V.$$

The set $\{u_p(g) : p \in [\![1, d-1]\!]\}$ is an *arbitrary* orthogonal choice of the axes (ordered in decreasing length) of the ellipsoid $\{Av : \|v\| = 1\}$, and, by construction, for every $v \in g^{-1}u_p(g)$ one has $\|gv\| = \sigma_p(g)\|v\|$. Let

$$U_p(g) = u_1(g) \oplus \cdots \oplus u_p(g) = k_g \cdot (e_1 \oplus \cdots \oplus e_p).$$

If g is such that $\sigma_p(g) > \sigma_{p+1}(g)$, then we say that g *has a gap of index* p. In that case the decomposition

$$U_{d-p}(g^{-1}) \oplus g^{-1}(U_p(g))$$

is orthogonal (see [39, Remark 2.4]) and, if $\mathbb{K}$ is Archimedean, the p–dimensional space $U_p(g)$ is independent of the Cartan decomposition of g.

[3] Recall that for $\mathbb{K}$ non-Archimedean a decomposition $V = V_1 \oplus \cdots \oplus V_k$ is *orthogonal* if $\|\sum v_i\| = \max_i \|v_i\|$ for every $v_i \in V_i$.

We will denote by $\Pi = \{\mathsf{a}_1, \dots, \mathsf{a}_{d-1}\}$ the root system of $\mathsf{PGL}(V_{\mathbb{K}})$, and, given a subset $\theta \subset \Pi$, by $\mathcal{F}_\theta$ the associated partial flag manifold. Given $\theta \subset \Pi$ we also denote by $U^\theta(g)$ the partial flag $U^\theta(g) = \{U_p(g) : \mathsf{a}_p \in \theta\}$. The *$\theta$–basin of attraction of g*

$$B_{\theta,\alpha}(g) = \left\{x^\theta \in \mathcal{F}_\theta(\mathbb{K}^d) : \min_{\mathsf{a}_p \in \theta} \angle(x^p, U_{d-p}(g^{-1})) > \alpha\right\} \tag{2-1}$$

is the complement of the α–neighborhood of $U^{\theta^c}(g^{-1})$. When θ consists of a single root a, we will write $B_{\mathsf{a},\alpha}(g)$ instead of $B_{\{\mathsf{a}\},\alpha}(g)$.

Remark 2.1 If g has a gap of index p, then $U_{d-p}(g^{-1})$ is well defined if $\mathbb{K}$ is Archimedean, and any two possible choices have distance at most $(\sigma_{p+1}/\sigma_p)(g)$ if $\mathbb{K}$ is non-Archimedean. It follows that, also in the non-Archimedean case, $B_{\theta,\alpha}(g)$ only depends on K provided α is bigger than the minimal singular value gap.

We recall for later use the following lemma, which explains the choice of the term basin of attraction:

Lemma 2.2 (Bochi, Potrie and Sambarino [6, Lemma A.6]) *For every $g \in \mathsf{PGL}_d(\mathbb{K})$ and $x \in B_{\mathsf{a}_1,\alpha}(g)$,*

$$d(U_1(g), g \cdot x) \le \frac{1}{\sin(\alpha)} \frac{\sigma_2}{\sigma_1}(g).$$

2.2 Anosov representations

Let Γ be a word-hyperbolic group with identity element e, and fix a finite symmetric generating set S_Γ. For $\gamma \in \Gamma \setminus \{e\}$ denote by $|\gamma|$ the least number of elements of S_Γ needed to write γ as a word on S, and define the induced distance $d_\Gamma(\gamma, \eta) = |\gamma^{-1}\eta|$. A geodesic segment on Γ is a sequence $\{\alpha_i\}_0^k$ of elements in Γ such that $d_\Gamma(\alpha_i, \alpha_j) = |i-j|$.

Definition 2.3 A representation $\rho\colon \Gamma \to \mathsf{PGL}_d(\mathbb{K})$ is a_p–*Anosov*[4] if there exist positive constants c and μ, the a_p–*Anosov constants of* ρ, such that for all $\gamma \in \Gamma$,

$$\frac{\sigma_{p+1}}{\sigma_p}(\rho(\gamma)) \le c e^{-\mu|\gamma|}. \tag{2-2}$$

An a_1–Anosov representation will be called *projective Anosov*.

The following result was proven in Bochi, Potrie and Sambarino [6] for $\mathbb{K} = \mathbb{R}$. The same arguments also give the result for any local field.

[4]In the language of Bochi, Potrie and Sambarino [6, Section 3.1], an a_p–Anosov representation is called *p–dominated*. It was proven by Kapovich, Leeb and Porti [33] that if a group Γ admits an Anosov representation, it is necessarily word hyperbolic. See also Bochi, Potrie and Sambarino [6] for a different approach.

Proposition 2.4 [6, Lemma 2.5] *Let $\rho\colon \Gamma \to \mathsf{PGL}_d(\mathbb{K})$ be a projective Anosov representation. Then there exists $\eta_\rho > 0$ and $L \in \mathbb{N}$ such that, for every geodesic segment $\{\alpha_i\}_0^k$ in Γ through e with $|\alpha_0|, |\alpha_k| \geq L$,*

$$\angle\big(U_1(\rho(\alpha_k)), U_{d-1}(\rho(\alpha_0))\big) > \eta_\rho.$$

Proposition 2.4 is a key ingredient in the construction of boundary maps:

Proposition 2.5 [6, Lemma 4.9] *Let $\rho\colon \Gamma \to \mathsf{PGL}_d(\mathbb{K})$ be projective Anosov and $(\alpha_i)_0^\infty \subset \Gamma$ a geodesic ray based at the identity converging to $x \in \partial\Gamma$. Then*

$$\xi_\rho^1(x) := \lim_{i\to\infty} U_1(\rho(\alpha_i)) \quad \text{and} \quad \xi_\rho^{d-1}(x) := \lim_{i\to\infty} U_{d-1}(\rho(\alpha_i))$$

exist, do not depend on the ray, and define continuous ρ–equivariant transverse maps $\xi_\rho^1\colon \partial\Gamma \to \mathbb{P}(\mathbb{K}^d)$ and $\xi_\rho^{d-1}\colon \partial\Gamma \to \mathbb{P}((\mathbb{K}^d)^*)$. Furthermore, there are positive constants C and μ depending only on ρ such that

$$d\big(U_1(\rho(\alpha_k)), \xi_\rho^1(x)\big) \leq Ce^{-\mu k}.$$

The next lemma, concerning properties of boundary maps, will be valuable in Section 3.1:

Lemma 2.6 (Bochi, Potrie and Sambarino [6, Lemma 3.9]) *Let $\rho\colon \Gamma \to \mathsf{PGL}_d(\mathbb{K})$ be projective Anosov. Then there exist constants $\nu \in (0,1)$, $a_0 > 0$ and $a_1 > 0$ such that, for every $\gamma, \eta \in \Gamma$,*

$$d_\Gamma(\gamma, \eta) \geq \nu(|\gamma| + |\eta|) - a_0 - a_1\big|\log d\big(U_1(\rho(\gamma)), U_1(\rho(\eta))\big)\big|.$$

3 Hausdorff dimension of limit sets and the affinity exponent

Generalizing the definition given in Section 1, we define the affinity exponent h_ρ^{Aff} of a projective Anosov representation $\rho\colon \Gamma \to \mathsf{PGL}(V_\mathbb{K})$ as the critical exponent of the broken Dirichlet series

$$\Phi_\rho^{\mathsf{Aff}}(s) = \sum_{\gamma\in\Gamma} \left(\frac{\sigma_2}{\sigma_1}(\rho(\gamma)) \cdots \frac{\sigma_{p-1}}{\sigma_1}(\rho(\gamma))\right)^{d_\mathbb{K}} \left(\frac{\sigma_p}{\sigma_1}(\rho(\gamma))\right)^{s-d_\mathbb{K}(p-2)}$$

for $s \in [d_\mathbb{K}(p-2), d_\mathbb{K}(p-1)]$, where the dimension $d_\mathbb{K}$ of $\mathbb{P}^1(\mathbb{K})$ is 1 unless $\mathbb{K} = \mathbb{C}$ in which case $d_\mathbb{C} = 2$.

Recall furthermore that, for a metric space (Λ, d) and for $s > 0$, one defines its *s–capacity* as

$$\mathcal{H}^s(\Lambda) = \inf_\varepsilon \left\{ \sum_{U\in\mathfrak{U}} \operatorname{diam} U^s \;\middle|\; \mathfrak{U} \text{ is a covering of } \Lambda \text{ with } \sup_{U\in\mathfrak{U}} \operatorname{diam} U < \varepsilon \right\},$$

and that the *Hausdorff dimension* of Λ is defined by

$$\dim_{\mathrm{Hff}}(\Lambda) = \inf\{s : \mathcal{H}^s(\Lambda) = 0\} = \sup\{s : \mathcal{H}^s(\Lambda) = \infty\}. \tag{3-1}$$

The goal of the section is to prove:

Theorem 3.1 *Let $\mathbb{K}$ be a local field. If $\rho\colon \Gamma \to \mathsf{PGL}(V_{\mathbb{K}})$ is a_1–Anosov, then*

$$\dim_{\mathrm{Hff}}(\xi^1_\rho(\partial\Gamma)) \le h^{\mathrm{Aff}}_\rho.$$

The proof of Theorem 3.1 is elementary and based on the construction of a good cover of the image of the limit map (explicitly constructed in Section 3.1), which we show in Section 3.2 to be contained in ellipses of controlled axis.

3.1 Coarse cone types

In Pozzetti, Sambarino and Wienhard [39, Section 2.3.1] we used cone types at infinity to construct well-behaved coverings of the boundary of the group. We now introduce a coarse version of these sets, which will be more useful for our purposes.

Recall that a sequence $(\alpha_j)_0^\infty$ is a (c_0, c_1)–quasigeodesic if, for every pair j, l,

$$\frac{1}{c_0}|j-l| - c_1 \le d_\Gamma(\alpha_j, \alpha_l) \le c_0|j-l| + c_1.$$

We associate to every element γ a *coarse cone type at infinity*, consisting of endpoints at infinity of quasigeodesic rays based at γ^{-1} passing through the identity:

$$\mathcal{C}^{c_0,c_1}_\infty(\gamma) = \bigl\{[(\alpha_j)_0^\infty] \in \partial\Gamma \mid (\alpha_i)_0^\infty \text{ is a } (c_0,c_1)\text{–quasigeodesic}, \alpha_0 = \gamma^{-1}, e \in \{\alpha_j\}\bigr\}.$$

Hyperbolicity of Γ lets us understand the overlaps of coarse cone types; this will be crucial in Section 5.3 to guarantee bounded overlap of suitable covers of the limit set.

Proposition 3.2 *Let Γ be word hyperbolic. For every c_0 and c_1 there exists $C > 0$ such that if*

$$\gamma\mathcal{C}^{c_0,c_1}_\infty(\gamma) \cap \eta\mathcal{C}^{c_0,c_1}_\infty(\eta) \ne \varnothing$$

then

$$d_\Gamma(\gamma, \eta) \le \bigl||\gamma| - |\eta|\bigr| + C.$$

Proof Assume that $x \in \gamma\mathcal{C}^{c_0,c_1}_\infty(\gamma) \cap \eta\mathcal{C}^{c_0,c_1}_\infty(\eta)$. Since Γ is hyperbolic, by the Morse lemma there exists $K > 0$ (only depending on c_0, c_1 and the hyperbolicity constant of Γ) such that γ is at distance at most K from a geodesic ray from e to x. The same holds then for η, and, using the hyperbolicity of Γ again, we can assume up to enlarging the constant K (still depending on c_0 and c_1 only) that the two rays agree. This implies

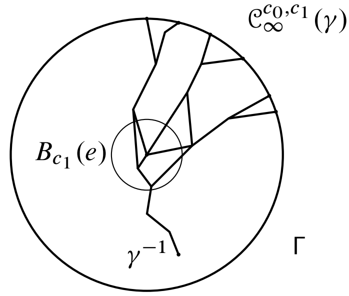

Figure 1: The coarse cone type at infinity. The black broken lines are (c_0, c_1)–quasigeodesics. All endpoints of geodesic rays from γ^{-1} intersecting the ball $B_{c_1}(e)$ clearly belong to $\mathcal{C}_\infty^{c_0,c_1}(\gamma)$.

that there exist g_0 and g_1 on a geodesic ray from e to x such that $d(\gamma, g_0) \le K$ and $d(\eta, g_1) \le K$. Since g_0 and g_1 lie in a geodesic $d(g_0, g_1) \le \big||g_0| - |g_1|\big|$, and thus

$$d(\gamma, \eta) \le 4K + \big||\gamma| - |\eta|\big|. \qquad \square$$

Our next goal is to show that, for an Anosov representation, the intersections of Cartan's basins of attraction $B_{\theta,\alpha}(\rho(\gamma))$ with the image of the boundary map are contained in the image of a suitably big coarse cone type of γ. Let $\theta \subset \Pi$ be a subset containing the first root a_1. We will denote by $\pi_{\theta,1}\colon \mathcal{F}_\theta(\mathbb{K}^d) \to \mathbb{P}(\mathbb{K}^d)$ the canonical projection. Recall from (2-1) that, for every α, we associate to each $g \in \mathsf{PGL}(V_{\mathbb{K}})$ a basin of attraction $B_{\theta,\alpha}(g) \subset \mathcal{F}_\theta$. We will now use Lemma 2.6 to show that, for every α, there exist c_0 and c_1 such that the intersection of a θ–basin of attraction $B_{\theta,\alpha}(\rho(\gamma))$ with the image of the boundary map is contained in a (c_0, c_1)–coarse cone type.

Proposition 3.3 *Let $\rho\colon \Gamma \to \mathsf{PGL}(V_{\mathbb{K}})$ be projective Anosov and consider $\alpha > 0$. There exist c_0 and c_1 only depending on α and ρ such that, for every $\theta \subset \Pi$ containing a_1 and every $\gamma \in \Gamma$,*

$$(\xi^1)^{-1}\big(\pi_{\theta,1}\big(B_{\theta,\alpha}(\rho(\gamma))\big)\big) \subset \mathcal{C}_\infty^{c_0,c_1}(\gamma).$$

Proof It is enough to show that if $\xi^1(x) \in \pi_{\theta,1}(B_{\theta,\alpha}(\rho(\gamma)))$ and $|\gamma|$ is big enough, then there is a quasigeodesic ray from γ^{-1} to x that passes through the identity and whose constants only depend on α and ρ. Consider a quasigeodesic ray $\{\alpha_j\}$ converging to x, and fix $1 > \alpha' > \alpha$. Since by assumption $\xi^1(x) \in B_{\mathsf{a}_1,\alpha}(\rho(\gamma))$, we can find a constant L depending only on ρ such that, for every $j > L$, it holds that $U_1(\rho(\alpha_i)) \in B_{\mathsf{a}_1,\alpha'}(\rho(\gamma))$. The uniformity of L follows from the last statement in

Proposition 2.5. By definition we have $\angle(U_1(\rho(\alpha_j)), U_{d-1}(\rho(\gamma^{-1}))) > \alpha'$, and thus, in particular, $d(U_1(\rho(\alpha_j)), U_1(\rho(\gamma^{-1})) > \alpha'$. Now let $(\alpha_j)_{i=0}^{-|\gamma|_S}$ be a geodesic segment with $\alpha_0 = e$ and $\alpha_{-|\gamma|_S} = \gamma$. Up to further enlarging α' and L (depending on the representation only), $d(U_1(\rho(\alpha_{-L})), U_1(\rho(\alpha_L)) > \alpha'$. Lemma 2.6 implies that the sequence $(\alpha_j)_{i=-|\gamma|_S}^{\infty}$, obtained as concatenation of the geodesic between γ^{-1} and the identity and the ray from the identity to x, is a quasigeodesic ray, thus the result. □

Corollary 3.4 *Let $\rho\colon \Gamma \to \mathsf{PGL}(V_{\mathbb{K}})$ be projective Anosov and consider $\alpha > 0$. There exists C only depending on α and ρ such that, for every $\theta \subset \Pi$ containing a_1, if*

$$\xi^1(\partial\Gamma) \cap \pi_{\theta,1}\big(\rho(\gamma) \cdot B_{\theta,\alpha}(\rho(\gamma)) \cap \rho(\eta) \cdot B_{\theta,\alpha}(\rho(\eta))\big) \neq \varnothing$$

then

$$d(\gamma, \eta) \leq \big||\gamma| - |\eta|\big| + C.$$

Proof This follows immediately by combining Propositions 3.3 and 3.2. □

In particular, we can use basins of attraction to construct coverings of the image of the boundary map with bounded overlap:

Proposition 3.5 (see Pozzetti, Sambarino and Wienhard [39, Lemma 2.22]) *Let $\rho\colon \Gamma \to \mathsf{PGL}(V_{\mathbb{K}})$ be projective Anosov. There exists α small enough such that, for every $T > 0$, the family of open sets*

$$\mathcal{U}_T := \{\rho(\gamma) \cdot B_{\mathsf{a}_1,\alpha}(\rho(\gamma)) : |\gamma| = T\}$$

defines an open covering of $\xi^1(\partial\Gamma)$. Furthermore there exists a constant C depending on α (and ρ) such that, for every $x \in \partial\Gamma$ and every T, $\xi(x)$ is contained in at most C elements of $\mathcal{U}_T$.

Proof Let $x \in \partial\Gamma$ and let $\{\gamma_j\}$ be a geodesic ray based at the identity representing x. Propositions 2.4 and 2.5 guarantee that there exists $\alpha = \alpha_\rho$ such that

$$\angle(\rho(\gamma_T^{-1})\xi^1(x), U_{d-1}(\rho(\gamma_T^{-1})) > \alpha;$$

therefore $\xi^1(x) \in \rho(\gamma_T) B_{\mathsf{a}_1,\alpha}(\rho(\gamma_T))$. The second statement is a direct consequence of Corollary 3.4. □

3.2 Ellipses

The purpose of this section is to prove that, for a projective Anosov representation, the set $\rho(\gamma) \cdot B_{\mathsf{a}_1,\alpha}(\rho(\gamma))$ is coarsely contained in an ellipsoid with axes of size

$$\frac{\sigma_2}{\sigma_1}(\rho(\gamma)), \dots, \frac{\sigma_d}{\sigma_1}(\rho(\gamma)).$$

Definition 3.6 Let V be a d–dimensional $\mathbb{K}$–vector space with $\mathbb{K}$–norm $\|\cdot\|$. Let

$$u_1 \oplus \cdots \oplus u_d$$

be a $\mathbb{K}$–orthogonal decomposition and let $v = \sum v_j u_j$ be the associated decomposition of $v \in V$ for suitable $v_j \in \mathbb{K}$. Choose positive real numbers $a_2 \geq \cdots \geq a_d \geq 1$. If $\mathbb{K}$ is Archimedean, an *ellipsoid* about $\mathbb{K}u_1$ is the projectivization of

$$\left\{ v \in V \;\middle|\; |v_1|^2 \geq \sum_2^d (a_j |v_j|)^2 \right\}$$

for some $a_i > 0$. If instead $\mathbb{K}$ is non-Archimedean, an *ellipsoid* about $\mathbb{K}u_1$ is the projectivization of

$$\left\{ v \in V : |v_1| \geq \max_{2 \leq i \leq d} (a_j |v_j|) \right\}.$$

The vector spaces $u_1 \oplus u_j$ are *the axes* of the ellipsoid, and the *size* of the axis $u_1 \oplus u_j$ is $1/a_j$. We need the following covering lemma:

Lemma 3.7 *Let E be an ellipsoid with axis of size $1 \geq \beta_2 \geq \cdots \geq \beta_d$. For every $p \in [\![2, d]\!]$, E can be covered by*

$$2^{2p} \left(\frac{\beta_2 \cdots \beta_{p-1}}{\beta_p^{p-2}} \right)^{d_{\mathbb{K}}}$$

balls of radius $\sqrt{d}\,\beta_p$.

Proof We consider the affine chart of $\mathbb{P}(V)$ corresponding to $u_1 = 1$. The ellipsoid E is contained in the product of the balls $\{|v_i| \leq \beta_i\} \subset \mathbb{K}$ (it agrees with such a product if $\mathbb{K}$ is non-Archimedean). If $\mathbb{K}$ is Archimedean, the ball $\{|v_j| \leq \beta_j\}$ is contained in the union of $\lceil \beta_j/\beta_p \rceil^{d_{\mathbb{K}}}$ balls of radius β_p. Since the product of d balls of radius β_p is contained in a ball of radius $\sqrt{d}\,\beta_p$, we obtain that E can be covered by

$$\left\lceil \frac{\beta_2}{\beta_p} \right\rceil^{d_{\mathbb{K}}} \cdots \left\lceil \frac{\beta_{p-1}}{\beta_p} \right\rceil^{d_{\mathbb{K}}}$$

balls of radius $\sqrt{d}\,\beta_p$.

If instead $\mathbb{K}$ is non-Archimedean, the ball $\{|v_j| \leq \beta_j\}$ can be decomposed into $q^{\lfloor \log_q(\beta_j/\beta_p) \rfloor}$ balls of radius β_p, and hence E can be covered with

$$q^{\lfloor \log_q(\beta_2/\beta_p) \rfloor} \cdots q^{\lfloor \log_q(\beta_{p-1}/\beta_p) \rfloor}$$

balls of radius β_p. □

Proposition 3.8 *Consider $\alpha > 0$. For $g \in \mathsf{PGL}(V_{\mathbb{K}})$, the image of the corresponding Cartan basin of attraction $g \cdot B_{\mathsf{a}_1,\alpha}(g)$ is contained in the ellipsoid about $U_1(g)$ with axes $u_1(g) \oplus u_j(g)$ of size*

$$\frac{1}{\sin\alpha}\frac{\sigma_j}{\sigma_1}(g).$$

Proof Assume first that $\mathbb{K}$ is Archimedean. By definition of $B_{\mathsf{a}_1,\alpha}(g)$, for every $v \in \mathbb{K}^d$ with $\mathbb{K} \cdot v \in B_{\mathsf{a}_1,\alpha}(g)$,

$$|v_1|^2 \geq (\sin\alpha)^2 \sum_1^d |v_j|^2,$$

where $(v_1, \dots, v_d)$ are the coefficients in the decomposition of v with respect to the orthogonal splitting $V = \bigoplus g^{-1}u_j(g)$.

Since the coefficients w_j of gv in the decomposition induced by the orthogonal decomposition $V = \bigoplus u_j(g)$ satisfy $|w_j| = \sigma_j(g)|v_j|$,

$$|w_1|^2 = \sigma_1(g)^2|v_1|^2 \geq \sigma_1(g)^2(\sin\alpha)^2 \sum_{j=2}^d |v_j|^2 = \sigma_1(g)^2(\sin\alpha)^2 \sum_{j=2}^d \frac{1}{\sigma_j(g)^2}|w_j|^2.$$

One concludes that gv lies on the corresponding ellipsoid. The non-Archimedean case follows analogously. □

3.3 The lower bound on the affinity exponent

We now have all the ingredients needed to prove Theorem 3.1:

Proof For each $T > 0$, denote by $\mathcal{U}_T$ the covering of $\xi^1(\partial\Gamma)$ given by Proposition 3.5. By definition, $U = U_\gamma \in \mathcal{U}_T$ is of the form $\rho(\gamma) \cdot B_{\mathsf{a}_1,\alpha}(\rho(\gamma))$ for some γ satisfying $|\gamma| = T$. Proposition 3.8 applied to $\rho(\gamma)$ implies that $\rho(\gamma) \cdot B_{\mathsf{a}_1,\alpha}(\rho(\gamma))$ is contained in an ellipsoid about $\mathbb{K}u_1(\rho(\gamma))$ with axes of sizes

$$\frac{1}{\sin\alpha}\frac{\sigma_2}{\sigma_1}(\rho(\gamma)), \dots, \frac{1}{\sin\alpha}\frac{\sigma_d}{\sigma_1}(\rho(\gamma)).$$

Furthermore, since ρ is Anosov, we deduce from Lemma 2.2 that $\sup_{U \in \mathcal{U}_T} \operatorname{diam} U$ is arbitrarily small as T goes to infinity. Recall that the s–capacity $\mathcal{H}^s$ was defined by (3-1). Applying Lemma 3.7 to these ellipses and any $p \in [\![2, d]\!]$, we obtain

$$\mathcal{H}^s(\xi(\partial\Gamma)) \leq 2^{2p}\left(\frac{\sqrt{d}}{\sin\alpha}\right)^s \inf_T \sum_{|\gamma| \geq T} \left(\frac{\sigma_2}{\sigma_1}(\rho(\gamma)) \cdots \frac{\sigma_{p-1}}{\sigma_1}(\rho(\gamma))\right)^{d_{\mathbb{K}}} \left(\frac{\sigma_p}{\sigma_1}(\rho(\gamma))\right)^{s-d_{\mathbb{K}}(p-2)}.$$

By definition of the affinity exponent h_ρ^{Aff}, for all $s > h_\rho^{\mathsf{Aff}}$ the broken Dirichlet series

$$\sum_{\gamma\in\Gamma}\left(\frac{\sigma_2}{\sigma_1}(\rho(\gamma))\cdots\frac{\sigma_{p-1}}{\sigma_1}(\rho(\gamma))\right)^{d_{\mathbb{K}}}\left(\frac{\sigma_p}{\sigma_1}(\rho(\gamma))\right)^{s-d_{\mathbb{K}}(p-2)},\quad s\in[d_{\mathbb{K}}(p-2),d_{\mathbb{K}}(p-1)],$$

is convergent, and thus, for all $s > h_\rho^{\mathsf{Aff}}$,

$$2^p\left(\frac{\sqrt{d}}{\sin\alpha}\right)^s\inf_T\sum_{|\gamma|\geq T}\left(\frac{\sigma_2}{\sigma_1}(\rho(\gamma))\cdots\frac{\sigma_{p-1}}{\sigma_1}(\rho(\gamma))\right)^{d_{\mathbb{K}}}\left(\frac{\sigma_p}{\sigma_1}(\rho(\gamma))\right)^{s-d_{\mathbb{K}}(p-2)}=0.$$

As a result we conclude that for all $s > h_\rho^{\mathsf{Aff}}$ the s–capacity $\mathcal{H}^s(\xi(\partial\Gamma))$ vanishes; hence,

$$h_\rho^{\mathsf{Aff}}\geq\dim_{\mathrm{Hff}}(\xi(\partial\Gamma)).\qquad\square$$

The following generalization of Corollary 1.1 is also immediate:

Corollary 3.9 *If $\rho\colon\Gamma\to\mathsf{PGL}(V_{\mathbb{K}})$ is projective Anosov and $\dim_{\mathrm{Hff}}(\xi(\partial\Gamma))\geq pd_{\mathbb{K}}$, then*

$$\dim_{\mathrm{Hff}}(\xi(\partial\Gamma))\leq ph_\rho(\mathcal{J}_p^u).$$

Proof Observe that, for every $s\in[d_{\mathbb{K}}p,d_{\mathbb{K}}(p+1)]$, the value of the broken Dirichlet series defining the affinity exponent

$$\Phi_\rho^{\mathsf{Aff}}(s)=\sum_{\gamma\in\Gamma}\left(\frac{\sigma_2}{\sigma_1}(\rho(\gamma))\cdots\frac{\sigma_{p+1}}{\sigma_1}(\rho(\gamma))\right)^{d_{\mathbb{K}}}\left(\frac{\sigma_{p+2}}{\sigma_1}(\rho(\gamma))\right)^{s-d_{\mathbb{K}}p}$$

is smaller than or equal to the value of the series associated to the p^{th} unstable Jacobian divided by p:

$$\Phi_\rho^{\mathcal{J}_p^u/p}(s)=\sum_{\gamma\in\Gamma}\left(\frac{\sigma_2}{\sigma_1}(\rho(\gamma))\cdots\frac{\sigma_{p+1}}{\sigma_1}(\rho(\gamma))\right)^{\frac{s}{p}}.$$

Indeed,

$$\begin{aligned}\left(\frac{\sigma_{p+2}}{\sigma_1}(\rho(\gamma))\right)^{s-d_{\mathbb{K}}p}&=\left(\frac{\sigma_{p+2}}{\sigma_1}(\rho(\gamma))\right)^{p\left(\frac{s}{p}-d_{\mathbb{K}}\right)}\\&\leq\left(\frac{\sigma_2}{\sigma_1}(\rho(\gamma))\right)^{\frac{s}{p}-d_{\mathbb{K}}}\cdots\left(\frac{\sigma_{p+1}}{\sigma_1}(\rho(\gamma))\right)^{\frac{s}{p}-d_{\mathbb{K}}}.\end{aligned}$$

As a result, if $d_{\mathbb{K}}p\leq h_\rho^{\mathsf{Aff}}\leq d_{\mathbb{K}}(p+1)$, then $ph_\rho(\mathcal{J}_p^u)\geq h_\rho^{\mathsf{Aff}}$.

The result follows as, for all $k\in[\![1,d-1]\!]$ and $v\in\mathsf{E}^+$,

$$\frac{\mathcal{J}_{k-1}^u(v)}{k-1}\leq\frac{\mathcal{J}_k^u(v)}{k},$$

which implies $kh_\rho(\mathcal{J}_k^u)\leq(k-1)h_\rho(\mathcal{J}_{k-1}^u)$. $\square$

4 Semisimple algebraic groups

Let G be a connected semisimple $\mathbb{K}$–group, $\mathsf{G}_{\mathbb{K}}$ the group of its $\mathbb{K}$–points, A a maximal $\mathbb{K}$–split torus and $\boldsymbol{X}(\mathsf{A})$ the group of its $\mathbb{K}^*$–characters. Consider the real vector space $\mathsf{E}^* = \boldsymbol{X}(\mathsf{A}) \otimes_{\mathbb{Z}} \mathbb{R}$ and E its dual. For every $\chi \in \boldsymbol{X}(\mathsf{A})$, we denote by χ^ω the corresponding linear form on E.

4.1 Restricted roots and parabolic groups

Let Σ be the set of restricted roots of A in $\mathfrak{g}$. Then the set Σ^ω is a root system of E^*. Let Σ^+ be a system of positive roots and Π the associated subset of simple roots. Let E^+ be the Weyl chamber determined by the positive roots $(\Sigma^\omega)^+$.

Let W be the Weyl group of Σ. It is isomorphic to the quotient of the normalizer $N_{\mathsf{G}_{\mathbb{K}}}(\mathsf{A}_{\mathbb{K}})$ of $\mathsf{A}_{\mathbb{K}}$ in $\mathsf{G}_{\mathbb{K}}$ by its centralizer $Z_{\mathsf{G}_{\mathbb{K}}}(\mathsf{A}_{\mathbb{K}})$. Let $\mathrm{i}\colon \mathsf{E} \to \mathsf{E}$ be the opposition involution: if $u\colon \mathsf{E} \to \mathsf{E}$ is the unique element in the Weyl group with $u(\mathsf{E}^+) = -\mathsf{E}^+$, then $\mathrm{i} = -u$.

A subset $\Theta \subset \Pi$ determines a pair of opposite parabolic subgroups P_Θ and $\check{\mathsf{P}}_\Theta$ whose Lie algebras are defined by

$$\mathfrak{p}_\Theta = \bigoplus_{\mathsf{a} \in \Sigma^+ \cup \{0\}} \mathfrak{g}_{\mathsf{a}} \oplus \bigoplus_{\mathsf{a} \in \langle \Pi - \Theta \rangle} \mathfrak{g}_{-\mathsf{a}} \quad \text{and} \quad \check{\mathfrak{p}}_\Theta = \bigoplus_{\mathsf{a} \in \Sigma^+ \cup \{0\}} \mathfrak{g}_{-\mathsf{a}} \oplus \bigoplus_{\mathsf{a} \in \langle \Pi - \Theta \rangle} \mathfrak{g}_{\mathsf{a}}.$$

The group $\check{\mathsf{P}}_\Theta$ is conjugate to the parabolic group $\mathsf{P}_{\mathrm{i}\Theta}$. Let

$$\mathfrak{l}_\Theta = \mathfrak{p}_\Theta \cap \check{\mathfrak{p}}_\Theta$$

be the Lie algebra of the associated Levi group.

The $\mathbb{K}$–*flag space* associated to Θ is $\mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}}) = \mathsf{G}_{\mathbb{K}}/\mathsf{P}_{\Theta,\mathbb{K}}$, and the $\mathsf{G}_{\mathbb{K}}$ orbit of the pair $([\mathsf{P}_{\Theta,\mathbb{K}}], [\check{\mathsf{P}}_{\Theta,\mathbb{K}}])$ is the unique open orbit for the action of $\mathsf{G}_{\mathbb{K}}$ in the product $\mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}}) \times \mathcal{F}_{\mathrm{i}\Theta}(\mathsf{G}_{\mathbb{K}})$. This orbit is denoted by $\mathcal{F}^{(2)}_\Theta(\mathsf{G}_{\mathbb{K}})$.

For $y \in \mathcal{F}_{\mathrm{i}\Theta}(\mathsf{G}_{\mathbb{K}})$ denote by

$$\operatorname{Ann}(y) = \{x \in \mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}}) : (x, y) \notin \mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}})^{(2)}\} \tag{4-1}$$

the closed submanifold of flags in $\mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}})$ that are not transverse to y.

Denote by $(\cdot\,,\cdot)$ a W–invariant inner product on E and also the induced inner product on E^*, define

$$\langle \chi, \psi \rangle = \frac{2(\chi, \psi)}{(\psi, \psi)},$$

and let $\{\omega_{\mathsf{a}}\}_{\mathsf{a}\in\Pi}$ be the *dual basis* of Π, ie $\langle \omega_{\mathsf{a}}, \mathsf{b}\rangle = d_{\mathsf{a}}\delta_{\mathsf{ab}}$, where $d_{\mathsf{a}} = 1$ if $2\mathsf{a} \notin (\Sigma^\omega)^+$ and $d_{\mathsf{a}} = 2$ otherwise. The linear form ω_{a} is *the fundamental weight* associated to a.

4.2 Cartan decomposition

Let $\nu\colon \mathsf{A}_{\mathbb{K}} \to \mathsf{E}$ be defined, for $z \in \mathsf{A}_{\mathbb{K}}$, as the unique vector in E such that for every $\chi \in \boldsymbol{X}(\mathsf{A})$,

$$\chi^\omega(\nu(z)) = \log|\chi(z)|.$$

Define $\mathsf{A}_{\mathbb{K}}^+ = \nu^{-1}(\mathsf{E}^+)$.

Let $K \subset \mathsf{G}_{\mathbb{K}}$ be a compact group that contains a representative for every element of the Weyl group W; that is, $N_{\mathsf{G}_{\mathbb{K}}}(\mathsf{A}_{\mathbb{K}}) = (N_{\mathsf{G}_{\mathbb{K}}}(\mathsf{A}_{\mathbb{K}}) \cap K)\mathsf{A}_{\mathbb{K}}$, where $N_{\mathsf{G}_{\mathbb{K}}}$ is the normalizer. One has $\mathsf{G}_{\mathbb{K}} = K\mathsf{A}_{\mathbb{K}}^+K$, and if $z, w \in \mathsf{A}_{\mathbb{K}}^+$ are such that $z \in KwK$, then $\nu(z) = \nu(w)$. There exists thus a function

$$a\colon \mathsf{G}_{\mathbb{K}} \to \mathsf{E}^+,$$

such that for every $g_1, g_2 \in \mathsf{G}_{\mathbb{K}}$ one has that $g_1 \in Kg_2K$ if and only if $a(g_1) = a(g_2)$. It is called the *Cartan projection* of $\mathsf{G}_{\mathbb{K}}$.

In the case of $G_{\mathbb{K}} = \mathsf{PGL}(V_{\mathbb{K}})$ this is nothing but the ordered list of semihomothecy ratios defined in Section 2.1.

4.3 Representations of $\mathsf{G}_{\mathbb{K}}$

Let $\Lambda\colon \mathsf{G} \to \mathsf{PGL}(V)$ be a finite-dimensional irreducible representation that is also a rational map between algebraic varieties and denote by $\phi_\Lambda\colon \mathfrak{g} \to \mathfrak{sl}(V)$ the Lie algebra homomorphism associated to Λ. Then the *weight space* associated to $\chi \in \boldsymbol{X}(\mathsf{A})$ is the vector space

$$V_\chi = \{v \in V : \phi_\Lambda(a)v = \chi(a)v \text{ for all } a \in \mathsf{A}_{\mathbb{K}}\},$$

and if $V_\chi \neq 0$ then we say that $\chi^\omega \in \mathsf{E}^*$ is a *restricted weight* of Λ. Theorem 7.2 of Tits [47] states that the set of weights has a unique maximal element with respect to the order $\chi \geq \psi$ if $\chi - \psi$ is positive on E^+. This is called *the highest weight* of Λ and denoted by χ_Λ.

Definition 4.1 Let Θ_Λ be the set of simple roots $\mathsf{a} \in \Pi$ such that $\chi_\Lambda - \mathsf{a}$ is still a weight of Λ.

Remark 4.2 The subset Θ_Λ is the subset of simple roots such that, for $\mathsf{a} \in \Sigma^+$, $n \in \mathfrak{g}_{-\mathsf{a}}$ and $v \in \chi_\Lambda$, we have $\phi_\Lambda(n)v = 0$ if and only if $\mathsf{a} \in \langle \Pi - \Theta_\Lambda \rangle$.

Definition 4.3 We denote by $\|\cdot\|_\Lambda$ a good norm on V, invariant under ΛK, and such that $\Lambda \mathsf{A}_{\mathbb{K}}$ consists of semihomothecies; if $\mathbb{K}$ is Archimedean the existence of such a norm is classical, and if $\mathbb{K}$ is non-Archimedean then this is the content of Quint [40, théorème 6.1].

For every $g \in \mathsf{G}_{\mathbb{K}}$,

$$\log \|\Lambda g\|_\Lambda = \chi_\Lambda(a(g)). \tag{4-2}$$

If $g = k_g z_g l_g$ with $k, l \in K$ and $z_g \in \mathsf{A}_{\mathbb{K}}^+$, then for all $v \in \Lambda(l_g^{-1})V_{\chi_\Lambda}$ one has $\|\Lambda g(v)\|_\Lambda = \|\Lambda g\|_\Lambda \|v\|_\Lambda$.

Denote by W_{χ_Λ} the $\Lambda \mathsf{A}_{\mathbb{K}}$–invariant complement of V_{χ_Λ}. Note that the stabilizer in $\mathsf{G}_{\mathbb{K}}$ of W_{χ_Λ} is $\check{\mathsf{P}}_{\Theta,\mathbb{K}}$, and thus one has a map of flag spaces

$$(\xi_\Lambda, \xi_\Lambda^*)\colon \mathcal{F}^{(2)}_{\Theta_\Lambda}(\mathsf{G}_{\mathbb{K}}) \to \mathcal{G}^{(2)}_{\dim V_{\chi_\Lambda}}(V), \tag{4-3}$$

a proper embedding which is a homeomorphism onto its image. Here $\mathcal{G}^{(2)}_{\dim V_{\chi_\Lambda}}(V)$ is the open $\mathsf{PGL}(V_{\mathbb{K}})$–orbit in the product of the Grassmannian of $(\dim V_{\chi_\Lambda})$–dimensional subspaces and the Grassmannian of $(\dim V - \dim V_{\chi_\Lambda})$–dimensional subspaces.

Proposition 4.4 (Tits [47]; see also Humphreys [30, Chapter XI]) *For each* $\mathsf{a} \in \Pi$ *there exists a finite-dimensional rational irreducible representation* $\Lambda_{\mathsf{a}}\colon \mathsf{G} \to \mathsf{PSL}(V_{\mathsf{a}})$ *such that* $\chi_{\Lambda_{\mathsf{a}}}$ *is an integer multiple of the fundamental weight* ω_{a} *and* $\dim V_{\chi_{\Lambda_{\mathsf{a}}}} = 1$. *All other weights of* Λ_{a} *are of the form*

$$\chi_{\mathsf{a}} - \mathsf{a} - \sum_{\mathsf{b} \in \Pi} n_{\mathsf{b}} \mathsf{b},$$

where $n_{\mathsf{b}} \in \mathbb{N}$.

We will fix from now on such a set of representations and call them, for each $\mathsf{a} \in \Pi$, the *Tits representation associated to* a.

4.4 The center of the Levi group $\mathsf{P}_{\Theta,\mathbb{K}} \cap \check{\mathsf{P}}_{\Theta,\mathbb{K}}$

We now consider the vector subspace

$$\mathsf{E}_\Theta = \bigcap_{\mathsf{a} \in \Pi - \Theta} \ker \mathsf{a}^\omega,$$

together with the unique projection $\pi_\Theta\colon \mathsf{E} \to \mathsf{E}_\Theta$ that is invariant under the subgroup W_Θ of the Weyl group spanned by reflections associated to roots in $\Pi - \Theta$:

$$W_\Theta = \{w \in W : w(v) = v \text{ for all } v \in \mathsf{E}_\Theta\}.$$

The dual space $(\mathsf{E}_\Theta)^*$ is canonically the subspace of E^* of π_Θ–invariant linear forms and it is spanned by the fundamental weights of roots in Θ:

$$(\mathsf{E}_\Theta)^* = \{\varphi \in \mathsf{E}^* : \varphi \circ \pi_\Theta = \varphi\} = \langle \omega_{\mathsf{a}} : \mathsf{a} \in \Theta \rangle.$$

Since $\pi_\Theta^2 = \pi_\Theta$, precomposition with π_Θ induces a projection $\mathsf{E}^* \to (\mathsf{E}_\Theta)^*$ denoted by

$$\varphi \mapsto \varphi^\Theta := \varphi \circ \pi_\Theta.$$

Examples 4.5 and 4.6 will be relevant in Sections 7 and 8, respectively.

Example 4.5 Let $\mathsf{G}_{\mathbb{K}} = \mathsf{PGL}(V_{\mathbb{K}})$ and, as above, denote by $\mathsf{a}_k \in \mathsf{E}^*$ the k^{th} simple root, so that $\mathsf{a}_k(a_1, \dots, a_d) = a_k - a_{k+1}$. We then choose $p \in [\![2, d-2]\!]$ and let $\Theta = \{\mathsf{a}_1, \mathsf{a}_p, \mathsf{a}_{d-1}\}$, so that

$$\mathsf{E}_\Theta = \{(a_1, \dots, a_d) \in \mathsf{E} : a_2 = \cdots = a_p \text{ and } a_{p+1} = \cdots = a_{d-1}\}$$

is three-dimensional. Using the fact that the fundamental weights ω_i (for $i = 1, p, d-1$) belong to $(\mathsf{E}_\Theta)^*$, one checks that the projection is

$$\begin{aligned}
&\varepsilon_1(\pi_\Theta(a)) = a_1,\\
&\varepsilon_i(\pi_\Theta(a)) = \frac{a_2 + \cdots + a_p}{p-1} = \frac{\omega_p - \omega_1}{p-1}(a) && \text{for every } i \in [\![2, p]\!],\\
&\varepsilon_i(\pi_\Theta(a)) = \frac{a_{p+1} + \cdots + a_{d-1}}{d-p-1} = \frac{\omega_{d-1} - \omega_p}{d-p-1}(a) && \text{for every } i \in [\![p+1, d-1]\!],\\
&\varepsilon_d(\pi_\Theta(a)) = a_d.
\end{aligned}$$

Then

$$\mathsf{a}_p^\Theta = \frac{\omega_p - \omega_1}{p-1} - \frac{\omega_{d-1} - \omega_p}{d-p-1}$$

and $\mathsf{a}_p^\Theta|_{\mathsf{E}^+\setminus\{0\}} \geq \mathsf{a}_p|_{\mathsf{E}^+\setminus\{0\}}$.

Example 4.6 Consider the group $\mathsf{SO}(p,q)$ of transformations in $\mathsf{PSL}_{p+q}(\mathbb{R})$ preserving a signature (p,q) bilinear form with $p < q$. One has

$$\mathsf{E} = \{(a_1, \dots, a_p) : a_i \in \mathbb{R}\}$$

equipped with the root system

$$\Sigma^\omega = \{\varepsilon_i : i \in [\![1, p]\!]\} \cup \{a \mapsto a_i - a_j : i, j \in [\![1, p]\!]\}.$$

A Weyl chamber can be chosen as

$$\mathsf{E}^+ = \{a \in \mathsf{E} : a_i \geq a_{i+1} \text{ for all } i \in [\![1, p-1]\!] \text{ and } a_p \geq 0\},$$

with the associated set of simple roots

$$\Pi = \{\mathsf{a}_i : i \in [\![1, p-1]\!]\} \cup \{\varepsilon_p\}.$$

Consider then $\Theta = \{\mathsf{a}_i : i \in [\![1, p-1]\!]\}$, so that $\mathsf{E}_\Theta = \ker \varepsilon_p$ and thus $\mathsf{a}_i \in (\mathsf{E}_\Theta)^*$ for $i \in [\![1, p-2]\!]$. Moreover,

$$\mathsf{a}^\Theta_{p-1} = \varepsilon_{p-1},$$

and one has that $\mathsf{a}^\Theta_{p-1}|_{\mathsf{E}^+\setminus\{0\}} \geq \mathsf{a}_{p-1}|_{\mathsf{E}^+\setminus\{0\}}$.

4.5 Gromov product

Recall from Sambarino [45] that the *Gromov product*[5] based at K is the map

$$(\cdot|\cdot)_K : \mathcal{F}^{(2)}_\Theta(\mathsf{G}_\mathbb{K}) \to \mathsf{E}_\Theta,$$

defined to be the unique vector $(x|y)_K \in \mathsf{E}_\Theta$ such that

$$\chi_\mathsf{a}((x|y)_K) = -\log \sin \angle_{\|\|_{\Lambda_\mathsf{a}}}(\xi_{\Lambda_\mathsf{a}} x, \xi^*_{\Lambda_\mathsf{a}} y)$$

for all $\mathsf{a} \in \Theta$, where χ_a is the fundamental weight associated to the Tits representation Λ_a of a. Note that

$$\max_{\mathsf{a}\in\Theta} \chi_\mathsf{a}((x|y)_K) = \max_{\mathsf{a}\in\Theta} \big|\chi_\mathsf{a}((x|y)_K)\big| = -\log \min_{\mathsf{a}\in\Theta} \sin \angle_{\|\|_{\Lambda_\mathsf{a}}}(\xi_{\Lambda_\mathsf{a}} x, \xi^*_{\Lambda_\mathsf{a}} y). \tag{4-4}$$

One has the following remark from Bochi, Potrie and Sambarino [6]:

Remark 4.7 [6, Remark 8.11] Let $\Lambda : \mathsf{G} \to \mathsf{PGL}(V)$ be a finite-dimensional rational irreducible representation. If $(x, y) \in \mathcal{F}^{(2)}_{\Theta_\Lambda}(\mathsf{G}_\mathbb{K})$ then

$$(\xi_\Lambda x | \xi^*_\Lambda y)_{\|\|_\Lambda} = \chi_\Lambda((x|y)_K),$$

where $\|\|_\Lambda$ denotes the (stabilizer of the) inner product on V such that ΛK is orthogonal (see Definition 4.3).

4.6 Iwasawa cocycle and its relation to representations of G

Another important decomposition of Lie groups that will play a role in our work is the Iwasawa decomposition

$$\mathsf{G}_\mathbb{K} = K\mathsf{A}_\mathbb{K}\mathsf{U}_{\Pi,\mathbb{K}},$$

[5] This is the negative of the product defined in [45].

where $\mathsf{P}_{\Pi,\mathbb{K}}$ is the minimal parabolic subgroup and $\mathsf{U}_{\Pi,\mathbb{K}}$ is its unipotent radical. For a general local field $\mathbb{K}$ the decomposition of an element is not necessarily unique, but if $z_1, z_2 \in \mathsf{A}_{\mathbb{K}}$ are such that $z_1 \in K z_2 \mathsf{U}_{\Pi,\mathbb{K}}$, then $\nu(z_1) = \nu(z_2)$.

Quint used the Iwasawa decomposition to define the Iwasawa cocycle

$$\mathrm{b}_{\Pi}(g, x) = \nu(z),$$

where $x = k[\mathsf{P}_{\Theta,\mathbb{K}}] \in \mathcal{F}_{\Theta}(\mathsf{G}_{\mathbb{K}})$ with $k \in K$ and $g \in \mathsf{G}_{\mathbb{K}}$, and gk has Iwasawa decomposition $gk = lzu$.

Lemma 4.8 (Quint [42, lemmes 6.1 et 6.2]) *The map $p_{\Theta} \circ \mathrm{b}_{\Pi}$ factors through a map $\mathrm{b}_{\Theta} \colon \mathsf{G}_{\mathbb{K}} \times \mathcal{F}_{\Theta}(\mathsf{G}_{\mathbb{K}}) \to \mathsf{E}_{\Theta}$. The map b_{Θ} verifies the cocycle relation: for every $g, h \in \mathsf{G}_{\mathbb{K}}$ and $x \in \mathcal{F}_{\Theta,\mathbb{K}}(\mathsf{G}_{\mathbb{K}})$,*

$$\mathrm{b}_{\Theta}(gh, x) = \mathrm{b}_{\Theta}(g, hx) + \mathrm{b}_{\Theta}(h, x).$$

One also has the following behavior of b_{Θ} under the representations of G:

Lemma 4.9 (Quint [42, lemme 6.4]) *Suppose $\Lambda \colon \mathsf{G} \to \mathsf{PGL}(V)$ is a proximal irreducible representation. Then for every $x \in \mathcal{F}_{\Theta_{\Lambda}}(\mathsf{G}_{\mathbb{K}})$ and $g \in \mathsf{G}_{\mathbb{K}}$,*

$$\chi_{\Lambda}(\mathrm{b}_{\Theta_{\Lambda}}(g, x)) = \log \frac{\|\Lambda(g)v\|_{\Lambda}}{\|v\|_{\Lambda}},$$

where $v \in \xi_{\Lambda}(x) \setminus \{0\}$.

4.7 Cartan attractors and Cartan's attracting basins

Consider $g \in \mathsf{G}_{\mathbb{K}}$ and let $g = k_g z_g l_g$ be a Cartan decomposition. Given $\Theta \subset \Pi$, the *Cartan attractor of g in* $\mathcal{F}_{\Theta}(\mathsf{G}_{\mathbb{K}})$ is defined by

$$U_{\Theta}(g) = U_{\Theta}^{K}(g) = k_g[\mathsf{P}_{\Theta,\mathbb{K}}],$$

and the *Cartan basin of g* is defined, for $\alpha > 0$, by

$$B_{\Theta,\alpha}(g) = \{x \in \mathcal{F}_{\Theta}(\mathsf{G}_{\mathbb{K}}) : (x \mid U_{\mathrm{i}\Theta}(g^{-1}))_K < \alpha\}.$$

Remark 4.10 If $\Lambda \colon \mathsf{G} \to \mathsf{PGL}(V)$ is a rational irreducible representation with $\Theta_{\Lambda} \subset \Theta$, then

$$\xi_{\Lambda}(U_{\Theta}(g)) = U_{\dim V_{\chi_{\Lambda}}}^{\|\|_{\Lambda}}(\Lambda(g)).$$

Notice that the flag $U_{\Theta}(g)$ is an arbitrary choice of a "most expanding" flag of type Θ for g. However, it is clear from the definition that given $\alpha > 0$ there exists a constant K_{α} such that, if $y \in \mathcal{F}_{\Theta}(\mathsf{G}_{\mathbb{K}})$ belongs to $B_{\Theta,\alpha}(g)$, then, for all $\mathsf{a} \in \Theta$,

$$|\chi_{\mathsf{a}}(a(g) - \mathrm{b}_{\Theta}(g, y))| \le K_{\alpha}. \tag{4-5}$$

4.8 The $\mathsf{PSL}_d(\mathbb{K})$ case

Given a good norm τ on $\mathbb{K}^d$, and considering the exterior power representations of $\mathsf{PSL}_d(\mathbb{K})$, one sees that Lemma 4.9 provides the following computation for the Iwasawa cocycle $\mathrm{b}\colon \mathsf{PSL}_d(\mathbb{K})\times\mathcal{F}(\mathbb{K}^d)\to\mathsf{E}$ associated to a maximal compact group stabilizing τ. For $p\in[\![1,d]\!]$ and given $g\in\mathsf{PSL}_d(\mathbb{K})$ and $x\in\mathcal{F}(\mathbb{K}^d)$,

$$\omega_p(\mathrm{b}(g,x)) = \log\frac{\|gv_1\wedge\cdots\wedge gv_p\|}{\|v_1\wedge\cdots\wedge v_p\|}, \tag{4-6}$$

where $\{v_1,\dots,v_p\}$ is any basis of the p–dimensional space x^p of x and $\|\cdot\|$ is the norm on $\bigwedge^p\mathbb{K}^d$ induced by τ.

Notice that, by definition, the number $\omega_p(\mathrm{b}(g,x))$ only depends on x^p, so in order to simplify notation we will also denote it by $\omega_p(\mathrm{b}(g,x^p))$.

5 Patterson–Sullivan measures in non-Anosov directions

An interesting quantity associated to a discrete subgroup $\Gamma<\mathsf{G}_\mathbb{K}$ is h_Γ^X, its critical exponent, which measures the exponential growth rate of orbit points in balls (in the symmetric space of $\mathsf{G}_\mathbb{K}$) as the radius grows. The theory of Quint's growth indicator function, which we briefly recall in Section 5.1, allows us to deduce information on h_Γ^X from information on the critical exponent of linear forms ϕ on the Weyl chamber E, which is often easier to handle with the aid of Patterson–Sullivan measures. When the discrete group $\Gamma<\mathsf{G}_\mathbb{K}$ is the image of an Anosov representation $\rho\colon\Gamma\to\mathsf{G}_\mathbb{K}$, and the form ϕ belongs to the dual of the Levi–Anosov subspace E_{θ_ρ}, the thermodynamical formalism applies (see Theorem 5.12).

In this section we will instead be interested in studying forms ϕ that do not belong to $(\mathsf{E}_{\theta_\rho})^*$. Our main result is Theorem 5.14, in which we show that, provided a representation ρ is Anosov with respect to some root, the existence of a Patterson–Sullivan measure in any flag manifold — and thus also in non-Anosov directions ϕ — has strong implications for the critical exponent of ϕ.

5.1 Quint's growth indicator

We recall here some definitions from Quint [41; 42].

Let $\Gamma\subset\mathsf{G}_\mathbb{K}$ be a discrete subgroup; its *Quint growth indicator function* [41]

$$\Psi_\Gamma\colon\mathsf{E}^+\to\mathbb{R}_+\cup\{-\infty\}$$

is defined as follows. Given a norm $\|\cdot\|$ on E and an open cone $\mathscr{C} \subset \mathsf{E}^+$, let $h_{\mathscr{C}}^{\|\|}$ be the critical exponent of the Dirichlet series

$$s \mapsto \sum_{\{g \in \Gamma : a(g) \in \mathscr{C}\}} e^{-s\|a(g)\|}$$

and define $\Psi_\Gamma \colon \mathsf{E}^+ \to \{-\infty\} \cup [0, \infty)$ by

$$\Psi_\Gamma(v) = \|v\| \inf_{v \in \mathscr{C}} h_{\mathscr{C}}^{\|\|},$$

where the infimum is taken over all open cones containing v. One can easily check that Ψ_Γ does not depend on the chosen norm $\|\cdot\|$ and is 1–positively homogenous.

Dually, one considers the growth on linear forms. The *limit* (*or Benoist* [4]) *cone* $\mathcal{L}_\Gamma$ of Γ is defined as the limit points of sequences $t_n a(g_n)$ where $(t_n)_{n \in \mathbb{N}} \subset \mathbb{R}_+$ converges to 0 and $(g_n)_{n \in \mathbb{N}} \subset \Gamma$. Denote its dual cone by

$$(\mathcal{L}_\Gamma)^* = \{\varphi \in \mathsf{E}^* : \varphi|_{\mathcal{L}_\Gamma \setminus \{0\}} \geq 0\},$$

and for $\varphi \in (\mathcal{L}_\Gamma)^*$ let $h_\Gamma(\varphi)$ be the critical exponent of the Dirichlet series

$$\sum_{g \in \Gamma} e^{-s\varphi(a(g))},$$

that is,

$$h_\Gamma(\varphi) = \limsup_{t \to \infty} \frac{1}{t} \log \#\{g \in \Gamma \mid \varphi(a(g)) < t\}.$$

Lemma 5.1 $\quad h_\Gamma(\min\{\phi_1, \dots, \phi_k\}) = \max\{h_\Gamma(\phi_1), \dots, h_\Gamma(\phi_k)\}.$

Proof One inequality is clear. For the other one,

$$\begin{aligned} h_\Gamma(\min\{\phi_1, \dots, \phi_k\}) &\leq \limsup_{t \to \infty} \frac{1}{t} \log \sum_{i=1}^{k} \#\{\gamma \in \mathsf{\Gamma} \mid \phi_i(a(\rho(\gamma))) < t\} \\ &\leq \limsup_{t \to \infty} \frac{1}{t} \log k \max_i \#\{\gamma \in \mathsf{\Gamma} \mid \phi_i(a(\rho(\gamma))) < t\} \\ &= \max\{h_\Gamma(\phi_1), \dots, h_\Gamma(\phi_k)\} \end{aligned}$$ □

One can then define the subset

$$\mathcal{D}_\Gamma = \{\varphi \in (\mathcal{L}_\Gamma)^* : h_\Gamma(\varphi) \in (0, 1]\}.$$

The next lemma is clear from the definitions, but is very useful in applications:

Lemma 5.2 *If ϕ belongs to $\mathcal{D}_\Gamma$, then $\phi + \psi \in \mathcal{D}_\Gamma$ for every $\psi \in (\mathcal{L}_\Gamma)^*$.*

The following result from Quint [41] allows one to deduce information on the critical exponent of various norms in terms of growth of linear functions, which are often easier to compute:

Proposition 5.3 (Quint [41]) *One has that*

$$\mathcal{D}_\Gamma = \{\varphi \in \mathsf{E}^* : \varphi(v) \geq \Psi_\Gamma(v) \text{ for all } v \in \mathsf{E}^+\},$$

and thus it is a convex set. Moreover, for any 1*–positively homogenous function* $\Theta\colon \mathsf{E}^+ \to \mathbb{R}$, *the critical exponent* $h_\Gamma(\Theta)$ *of the Dirichlet series*

$$s \mapsto \sum_{g \in \Gamma} e^{-s\Theta(a(g))}$$

can be computed as $h_\Gamma(\Theta) = \sup_{v \in \mathsf{E}^+} \Psi_\Gamma(v)/\Theta(v)$.

A useful property of the set $\mathcal{D}_\Gamma$ is provided by the next theorem.

Theorem 5.4 (Quint [41]) *If the Zariski closure of* Γ *is semisimple then* Ψ_Γ *is concave. Consequently, for every norm* $\|\cdot\|$ *on* E,

$$h_\Gamma^{\|\cdot\|} = \inf\{\|\varphi\|^* : \varphi \in \mathcal{D}_\Gamma\},$$

where $\|\cdot\|^*$ *is the induced operator norm on* E^*.

Remark 5.5 Recall that, if we endow the symmetric space (or the affine building) X associated to $\mathsf{G}_\mathbb{K}$ with a $\mathsf{G}_\mathbb{K}$–invariant Riemannian metric, there exists an Euclidean norm $\|\cdot\|_X$ on E such that, for every $g \in \mathsf{G}_\mathbb{K}$,

$$d_X([K], g[K]) = \|a(g)\|_X.$$

So Theorem 5.4 provides the following formula for the critical exponent of a discrete group with reductive Zariski closure in the symmetric space X:

$$h_\Gamma^X = \inf\{\|\phi\|_X^* : \phi \in \mathcal{D}_\Gamma\}.$$

The topological boundary $\mathcal{Q}_\Gamma$ of $\mathcal{D}_\Gamma$ will be called *Quint's indicator set of* Γ. We will also write

$$\mathcal{Q}_{\Gamma,\Theta} = \mathcal{Q}_\Gamma \cap (\mathsf{E}_\Theta)^*.$$

Let us record here a useful direct consequence of the convexity of $\mathcal{D}_\Gamma$:

Lemma 5.6 *Let* $\phi, \varphi \in (\mathcal{L}_\Gamma)^*$. *Then*

$$h_\Gamma(\phi + \varphi) \leq \frac{h_\Gamma(\phi) h_\Gamma(\varphi)}{h_\Gamma(\phi) + h_\Gamma(\varphi)}.$$

We end this subsection with a definition from Quint [42]:

Definition 5.7 Given $\Theta \subset \Pi$ and $\varphi \in (\mathsf{E}_\Theta)^*$, a (Γ, φ)–*Patterson–Sullivan measure* on $\mathcal{F}_\Theta(\mathsf{G}_\mathbb{K})$ is a finite Radon measure μ such that, for every $g \in \Gamma$,

$$\frac{\mathrm{d}g_*\mu}{\mathrm{d}\mu}(x) = e^{-\varphi(\mathrm{b}_\Theta(g^{-1},x))}.$$

5.2 Anosov representations with values in $\mathsf{G}_\mathbb{K}$

Let Γ be a discrete group and fix $\Theta \subset \Pi$.

Definition 5.8 A representation $\rho\colon \Gamma \to \mathsf{G}_\mathbb{K}$ is Θ–*Anosov* if there exist constants $c \geq 0$ and $\mu > 0$ such that, for every $\gamma \in \Gamma$ and $\mathsf{a} \in \Theta$,

$$\mathsf{a}(a(\rho(\gamma))) \geq \mu|\gamma| - c.$$

If $\rho\colon \Gamma \to \mathsf{G}_\mathbb{K}$ is Θ–Anosov and Λ_a is as in Proposition 4.4, then $\Lambda_\mathsf{a}\rho\colon \Gamma \to \mathsf{PGL}(V_\mathbb{K})$ is projective Anosov. In particular, Section 2.2 applies to arbitrary $\mathsf{G}_\mathbb{K}$ and one obtains the following result:

Theorem 5.9 (Kapovich, Leeb and Porti [34]) *If $\rho\colon \Gamma \to \mathsf{G}_\mathbb{K}$ is Θ–Anosov then Γ is word hyperbolic and there exist continuous equivariant maps $\xi_\rho^\Theta\colon \partial\Gamma \to \mathcal{F}_\Theta(\mathsf{G}_\mathbb{K})$ and $\xi_\rho^{\mathrm{i}\Theta}\colon \partial\Gamma \to \mathcal{F}_{\mathrm{i}\Theta}(\mathsf{G}_\mathbb{K})$ such that the product map $(\xi_\rho^\Theta, \xi_\rho^{\mathrm{i}\Theta})\colon \partial^{(2)}\Gamma \to \mathcal{F}_\Theta^{(2)}(\mathsf{G}_\mathbb{K})$ is transverse.*

We will sometime use the notation introduced in [39] and, if $x \in \partial\Gamma$ is a point, denote by

$$x_\rho^\Theta := \xi_\rho^\Theta(x) \in \mathcal{F}_\Theta(\mathsf{G}_\mathbb{K})$$

the image of x via the boundary map. If $\theta = \{\mathsf{a}_k\}$ consists of a single root, we will also write ξ_ρ^k and x_ρ^k instead of $\xi_\rho^{\{\mathsf{a}_k\}}$ and $x_\rho^{\{\mathsf{a}_k\}}$.

If $\Theta \subset \Pi$ contains the root a, we denote by $\pi_\mathsf{a}\colon \mathcal{F}_\theta(\mathsf{G}_\mathbb{K}) \to \mathcal{F}_\mathsf{a}(\mathsf{G}_\mathbb{K})$ the natural projection. It is easy to deduce from Corollary 3.4 the following more general statement:

Corollary 5.10 *Let $\rho\colon \Gamma \to \mathsf{G}_\mathbb{K}$ be a–Anosov and consider $\alpha > 0$. There exists C only depending on α and ρ such that, for every $\theta \subset \Pi$ containing a, if*

$$\xi_\rho^\mathsf{a}(\partial\Gamma) \cap \pi_\mathsf{a}\big(\rho(\gamma) \cdot B_{\theta,\alpha}(\rho(\gamma)) \cap \rho(\eta) \cdot B_{\theta,\alpha}(\rho(\eta))\big) \neq \varnothing$$

then

$$d(\gamma, \eta) \leq \big||\gamma| - |\eta|\big| + C.$$

Definition 5.11 Given a representation $\rho\colon \Gamma \to \mathsf{G}_{\mathbb{K}}$ we define its *Anosov–Levi space* as $(\mathsf{E}_{\Theta_\rho})^*$, where

$$\Theta_\rho = \{\mathsf{a} \in \Pi : \rho \text{ is } \mathsf{a}\text{–Anosov}\}.$$

It is spanned by the fundamental weights $\{\omega_{\mathsf{a}} : \mathsf{a} \in \Theta_\rho\}$.

A more precise description of the indicator set of ρ can be given on its Anosov–Levi space. The following is a combination of Bridgeman, Canary, Labourie and Sambarino [9, Theorem 1.3], Potrie and Sambarino [38, Proposition 4.11] and Sambarino [44]:

Theorem 5.12 *Let $\rho\colon \Gamma \to \mathsf{G}_{\mathbb{K}}$ be a representation. Then $\mathcal{Q}_{\rho(\Gamma),\Theta_\rho}$ is an analytic codimension-*1 *embedded submanifold of $(\mathsf{E}_{\Theta_\rho})^*$ that varies analytically with ρ. Moreover, its restriction to the dual of the vector space spanned by the periods is strictly convex.*

5.3 When some wall is not attained

The purpose of this subsection is to explore $\mathcal{Q}_{\rho(\Gamma)}$ in directions that are not controlled by the roots with respect to which ρ is Anosov.

Definition 5.13 Let $\rho\colon \Gamma \to \mathsf{G}_{\mathbb{K}}$ be an a–Anosov representation. Consider $\Theta \subset \Pi$ with $\mathsf{a} \in \Theta$ and let μ^φ be a $(\rho(\Gamma),\varphi)$–Patterson–Sullivan measure on $\mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}})$ for some $\varphi \in (\mathsf{E}_\Theta)^*$. We say that ρ is μ^φ*–irreducible* if, for every $y \in \mathcal{F}_{\mathrm{i}\Theta}(\mathsf{G}_{\mathbb{K}})$,

$$\mu^\varphi(\mathrm{Ann}(y)) < \mu^\varphi(\mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}})).$$

It is clear that if $\rho(\Gamma)$ is Zariski-dense in $\mathsf{G}_{\mathbb{K}}$ then it is μ^φ–irreducible for any Patterson–Sullivan measure. Even assuming Zariski-density, the following result is a refinement of Quint [42, théorème 8.1] when Θ contains a root with respect to which ρ is Anosov. Indeed, in the general case treated by Quint, one needs to control the mass of shadows on the flag space associated to $\Pi \setminus \Theta$, and, as a result, the existence of a $(\rho(\Gamma),\varphi)$–Patterson–Sullivan measure only ensures that $\varphi + \rho_{\theta^c}$ is in $\mathcal{D}_{\rho(\Gamma)}$, where ρ_{θ^c} is a suitable form that is nonnegative on the Weyl chamber. In our case, the Anosov condition with respect to one root in Θ permits us to control φ directly.

Theorem 5.14 *Let $\rho\colon \Gamma \to \mathsf{G}_{\mathbb{K}}$ be an a–Anosov representation. Consider $\Theta \subset \Pi$ with $\mathsf{a} \in \Theta$ and let μ^φ be a $(\rho(\Gamma),\varphi)$–Patterson–Sullivan measure on $\mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}})$ for some $\varphi \in (\mathsf{E}_\Theta)^*$. Assume ρ is μ^φ–irreducible, and that* $\operatorname{supp}\mu \subset \pi_{\mathsf{a}}^{-1}(\xi_\rho^{\mathsf{a}}(\partial\Gamma))$. *Then*

$$\varphi \in \mathcal{D}_{\rho(\Gamma)}.$$

The rest of the section is devoted to the proof of this result. We begin with the following lemma from Quint [42], who assumes that the representation is Zariski-dense, a hypothesis that is too strong for the applications we have in mind. We observe however that for the proof to work only μ^φ–irreducibility is needed. We sketch the proof for completeness.

Lemma 5.15 [42, lemme 8.2] *Let $\rho\colon \Gamma \to \mathsf{G}_{\mathbb{K}}$ be a representation and μ^φ be a $(\rho(\Gamma),\varphi)$–Patterson–Sullivan measure on $\mathcal{F}_\Theta(\mathsf{G}_{\mathbb{K}})$. Assume ρ is μ^φ–irreducible. Then there exists $\alpha_0 > 0$ such that, for every given $0 < \alpha < \alpha_0$, there exists $k > 0$, only depending on α, such that, for every $\gamma \in \Gamma$,*

$$k^{-1}e^{-\varphi(a(\rho(\gamma)))} \leq \mu^\varphi\big(\rho(\gamma)B_{\Theta,\alpha}(\rho(\gamma))\big) \leq ke^{-\varphi(a(\rho(\gamma)))}.$$

Proof Observe that μ^φ–irreducibility guarantees that there exist $\alpha, k > 0$ such that for every $\gamma \in \Gamma$ we have $\mu^\varphi(B_{\theta,\alpha}(\rho(\gamma))) \geq k$. Indeed, otherwise there would be a sequence of reals $\alpha_n \to 0$ and elements $\gamma_n \in \Gamma$ with $\mu^\varphi(B_{\theta,\alpha_n}(\rho(\gamma_n))) \leq 1/n$. We can assume, up to extracting a subsequence, that the complement of $B_{\theta,\alpha_n}(\rho(\gamma))$ converges to $\operatorname{Ann}(y)$ for some $y \in \mathcal{F}_{\mathrm{i}\theta}$, and this contradicts μ^φ–irreducibility.

The result then follows from the definition of the $(\rho(\Gamma),\phi)$–Patterson–Sullivan measure using (4-5). □

The rest of the proof of Theorem 5.14 is similar to the argument showing that if there exists a Patterson–Sullivan density of a given exponent, then this exponent must be greater than the critical exponent; see for example Sullivan [46] and Quint's notes [43, Theorem 4.11].

Proof of Theorem 5.14 We have to show that, for every $s > 0$,

$$\sum_{\gamma\in\Gamma} e^{-(1+s)\varphi(a(\rho(\gamma)))} < \infty.$$

Corollary 5.10 implies that given $\alpha > 0$ there exists $N \in \mathbb{N}$ such that, if $t > 0$ and

$$\Gamma_t = \{\gamma \in \Gamma : t \leq |\gamma| \leq t+1\},$$

then, for every $x \in \partial\Gamma$,

$$\#\{\gamma \in \Gamma_t : \pi_{\mathsf{a}}^{-1}(\xi_\rho^{\mathsf{a}}(x)) \cap \rho(\gamma)B_{\Theta,\alpha}(\rho(\gamma)) \neq \varnothing\} \leq N.$$

Lemma 5.15 now yields, for every $t \geq 0$,

$$\infty > \mu^\varphi(\pi_{\mathsf{a}}^{-1}(\xi_\rho^{\mathsf{a}}(\partial\Gamma))) \geq C \sum_{\gamma\in\Gamma_t} e^{-\varphi(a(\rho(\gamma)))}, \tag{5-1}$$

where C is independent of t. This is to say, there exists $K > 0$ independent of $t \in \mathbb{R}_+$ such that

$$\sum_{\gamma \in \Gamma_t} e^{-\varphi(a(\rho(\gamma)))} < K.$$

Since ρ is a–Anosov, for any norm N on E there exist positive δ and C such that

$$N(a(\rho(\gamma))) \geq \delta|\gamma| - C.$$

One concludes that, for every $s > 0$,

$$\sum_{\gamma \in \Gamma} e^{-\varphi(a(\rho(\gamma)) - sN(a(\rho(\gamma))} \leq \sum_{n=0}^{\infty} \sum_{\gamma \in \Gamma_n} e^{-\varphi(a(\rho(\gamma)))} e^{-sN(\rho(\gamma))} \leq K e^C \sum_{n=0}^{\infty} e^{-\delta s n} < \infty.$$

Consider now the counting measure ν on E defined by

$$\nu(B) = \#\{a(\Gamma) \cap B\}.$$

The above implies that the measure $\nu' = e^{-\varphi}\nu$ has growth indicator $\Psi_{\nu'} \leq 0$, and so [41, Corollary 3.1.5] gives

$$0 \geq \Psi_{e^{-\varphi}\nu} = \Psi_\nu - \varphi = \Psi_{\rho(\Gamma)} - \varphi,$$

as desired. □

6 Anosov representations with Lipschitz limit set

In this section we will prove Theorem A. We will hence fix some notation throughout this section.

Assumption 6.1 The group Γ will be a word-hyperbolic group whose boundary $\partial\Gamma$ is homeomorphic to a sphere of dimension d_Γ.[6] We will also fix a projective Anosov representation $\rho\colon \Gamma \to \mathsf{PSL}_d(\mathbb{R})$ such that the sphere $\xi^1_\rho(\partial\Gamma)$ is a Lipschitz submanifold of $\mathbb{P}(\mathbb{R}^d)$, ie it is locally the graph of a Lipschitz map. Note that we have restricted ourselves to $\mathbb{K} = \mathbb{R}$.

6.1 The p^{th} Jacobian

Given a line ℓ contained in a $(p+1)$–dimensional subspace V of $\mathbb{R}^d$, the space of infinitesimal deformations of ℓ inside V

$$\mathsf{T}_\ell\mathbb{P}(V) \subset \mathsf{T}_\ell\mathbb{P}(\mathbb{R}^d)$$

[6] It follows from [32, Theorem 4.4] that this is the case as soon as $\partial\Gamma$ has an open subset homeomorphic to $\mathbb{R}^{d_\Gamma}$.

carries a natural volume form induced by the choice of a scalar product τ on $\mathbb{R}^d$. Namely, if one considers the τ–orthogonal decomposition $V = \ell \oplus \ell_V^{\perp}$, then one canonically identifies $\mathsf{T}_\ell \mathbb{P}(V) = \hom(\ell, \ell_V^{\perp})$ and thus one can define $\Omega_{\ell,V} \in \bigwedge^p(\mathsf{T}_\ell \mathbb{P}(V))$ by

$$\Omega_{\ell,V}(\varphi_1, \dots, \varphi_p) = \frac{v \wedge \varphi_1(v) \wedge \cdots \wedge \varphi_p(v)}{\|v\|^{p+1}}$$

for any $v \in \ell \setminus \{0\}$.

Definition 6.2 The linear form $\mathcal{J}_p^u \in (\mathsf{E}_{\{\mathsf{a}_1,\mathsf{a}_{p+1}\}})^*$, defined by

$$\mathcal{J}_p^u = (p+1)\omega_1 - \omega_{p+1},$$

is called *the* p^{th} *unstable Jacobian.*

Lemma 6.3 *Given* $g \in \mathsf{PSL}_d(\mathbb{R})$ *and a partial flag* $(\ell, V) \in \mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_{p+1}\}}(\mathbb{R}^d)$*,*

$$g^*\Omega_{g\ell,gV} = \exp\bigl(-\mathcal{J}_p^u\bigl(\mathsf{b}_{\{\mathsf{a}_1,\mathsf{a}_{p+1}\}}(g, (\ell, V))\bigr)\bigr)\Omega_{\ell,V}.$$

Proof This is an explicit computation using (4-6) and the definition of $\Omega_{\ell,V}$.

Indeed, whenever $\varphi_1, \dots, \varphi_p \in \hom(\ell, \ell_V^{\perp})$ are linearly independent, for any $v \in \ell \setminus \{0\}$ the vectors $\{v, \varphi_1(v), \dots, \varphi_p(v)\}$ form a basis of V, and thus

$$\begin{aligned}
&g^*\Omega_{g\ell,gV}(\varphi_1, \dots, \varphi_p) \\
&\qquad = \Omega_{g\ell,gV}(g\varphi_1, \dots, g\varphi_p) \\
&\qquad = \frac{gv \wedge (g\varphi_1)(gv) \wedge \cdots \wedge (g\varphi_p)(gv)}{\|gv\|^{p+1}} = \frac{gv \wedge g(\varphi_1(v)) \wedge \cdots \wedge g(\varphi_p(v))}{\|gv\|^{p+1}} \\
&\qquad = \frac{gv \wedge g(\varphi_1(v)) \wedge \cdots \wedge g(\varphi_p(v))}{v \wedge \varphi_1(v) \wedge \cdots \wedge \varphi_p(v)} \frac{v \wedge \varphi_1(v) \wedge \cdots \wedge \varphi_p(v)}{\|v\|^{p+1}} \frac{\|v\|^{p+1}}{\|gv\|^{p+1}} \\
&\qquad = \exp\bigl(\omega_{p+1}(\mathsf{b}_{\{\mathsf{a}_1,\mathsf{a}_{p+1}\}}(g, V)) - (p+1)\omega_1(\mathsf{b}_{\{\mathsf{a}_1,\mathsf{a}_{p+1}\}}(g, \ell))\bigr)\Omega_{\ell,V}. \qquad \square
\end{aligned}$$

6.2 Existence of a $\mathcal{J}_{d_\Gamma}^u$–Patterson–Sullivan measure

Proposition 6.4 *Under Assumption 6.1, there exists a* $(\rho(\Gamma), \mathcal{J}_{d_\Gamma}^u)$*–Patterson–Sullivan measure on* $\mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_{d_\Gamma}\}}$*, which we will denote by* ν_ρ*.*

Proof It follows from Rademacher's theorem [20, Theorem 3.2] that $\xi_\rho^1(\partial\Gamma)$ has a well-defined Lebesgue measure class (see [22, Section 3.2]), and that Lebesgue almost every point $\xi_\rho^1(x) \in \xi_\rho^1(\partial\Gamma)$ has a well-defined tangent space. This defines a $(d_\Gamma+1)$–dimensional vector subspace $x_\rho^{d_\Gamma+1} \in \mathcal{F}_{\{\mathsf{a}_{d_\Gamma+1}\}}(\mathbb{R}^d)$ such that

$$\mathsf{T}_{\xi_\rho^1(x)}(\xi_\rho^1(\partial\Gamma)) = \hom(\xi_\rho^1(x), x_\rho^{d_\Gamma+1}/\xi_\rho^1(x)). \tag{6-1}$$

Consider the ρ–equivariant measurable map $\zeta_\rho\colon \xi^1_\rho(\partial\Gamma) \to \mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_{d_\Gamma+1}\}}(\mathbb{R}^d)$ defined by

$$\zeta_\rho(\xi^1_\rho(x)) = (\xi^1_\rho(x), x^{d_\Gamma+1}_\rho). \tag{6-2}$$

We can then define a volume form on $\xi^1_\rho(\partial\Gamma)$ via

$$\xi^1_\rho(x) \mapsto \Omega_{\zeta_\rho(\xi^1_\rho(x))}.$$

This form is defined Lebesgue almost everywhere and thus defines a Lebesgue measure on $\xi^1_\rho(\partial\Gamma)$, which we will denote by ν_ρ. Lemma 6.3 implies directly that the pushforward $(\zeta_\rho)_*\nu_\rho$ is the desired measure. □

6.3 When $\partial\Gamma$ is a circle

Recall from Section 1 that we say that ρ is *weakly irreducible* if the vector space $\operatorname{span}(\xi^1_\rho(\partial\Gamma))$ is the whole space.

Lemma 6.5 *Under Assumption 6.1 together with weakly irreducibility of ρ and $d_\Gamma = 1$, one has that ρ is μ^φ–irreducible for any $(\rho(\Gamma),\varphi)$–Patterson–Sullivan measure on $\mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_2\}}(\mathbb{R}^d)$ whose projection is absolutely continuous with the measure ν_ρ constructed in Proposition 6.4.*

Proof If this were not the case, there would exist $(W_0, P_0) \in \mathcal{F}_{\{\mathsf{a}_{d-2},\mathsf{a}_{d-1}\}}(\mathbb{R}^d)$ such that $\operatorname{Ann}(W_0, P_0)$ would have full μ^φ–mass; as ρ is projective Anosov we can furthermore assume that $P_0 = \xi^{d-1}_\rho(x)$ for some $x \in \partial\Gamma$ and thus the condition $\xi^1_\rho(y) \subset P_0$ only occurs for $y = x$.

Hence, since the projection of μ^φ is absolutely continuous with respect to ν_ρ, one has that, for μ^φ–almost every $\xi^1_\rho(x) \in \xi^1_\rho(\partial\Gamma)$, the vector space x^2_ρ from Section 6.2 intersects W_0.

Let us choose a scalar product τ on $\mathbb{R}^d$, and the induced distance function of $\mathbb{P}(\mathbb{R}^d)$. Let us denote by $[W_0]$ the quotient vector space $\mathbb{R}^d/W_0$. It is a 2–dimensional vector space and every line $\ell \notin W_0$ defines a line $[\ell \oplus W_0]$ in $[W_0]$. Moreover, for every $\delta > 0$, the double quotient projection

$$\pi : \{\ell \in \mathbb{P}(\mathbb{R}^d) : \angle_\tau(\ell, W_0) > \delta\} \to \mathbb{P}([W_0]),$$

defined by $\pi(\ell) = [[\ell \oplus W_0]]$, is Lipschitz.

We denote by $U_\delta \subset \xi^1_\rho(\partial\Gamma)$ the relative open subset defined by

$$U_\delta = \{\ell \in \xi^1_\rho(\partial\Gamma) : \angle_\tau(\ell, W_0) > \delta\}$$

and consider the Lipschitz map $\pi|_{U_\delta}\colon U_\delta \to \mathbb{P}([W_0])$. Since, by assumption, for μ^φ–almost every $\xi^1_\rho(x) \in \xi^1_\rho(\partial\Gamma)$ the plane x^2_ρ intersects W_0, one concludes from (6-1) that $\pi|_{U_\delta}$ has zero derivative ν_ρ–almost everywhere.

Since Lipschitz maps are absolutely continuous, and in particular satisfy the fundamental theorem of calculus, we deduce that $\pi|_{\xi^1_\rho(\partial\Gamma)}$ is constant. This implies that

$$\xi^1_\rho(\partial\Gamma) \subset W_0 \oplus \xi^1_\rho(x)$$

for any $x \in \partial\Gamma$, which contradicts the weak irreducibility assumption. □

We can now prove Theorem A when $d_\Gamma = 1$:

Corollary 6.6 *Let Γ be a word-hyperbolic group such that $\partial\Gamma$ is homeomorphic to a circle. Let $\rho\colon \Gamma \to \mathsf{PGL}_d(\mathbb{R})$ be a weakly irreducible a_1–Anosov representation such that $\xi^1_\rho(\partial\Gamma)$ is a Lipschitz curve. Then*

$$\mathsf{a}_1 \in \mathcal{Q}_{\rho(\Gamma)}.$$

Proof Note that $\mathsf{a}_1 = \mathcal{J}^u_1$ is the first unstable Jacobian. Since $\xi^1_\rho(\partial\Gamma)$ is a Lipschitz circle, it has Hausdorff dimension 1, and thus Corollary 1.1 implies that $h^{\mathsf{a}_1}_\rho \geq 1$.

On the other hand, Proposition 6.4 provides a $(\rho(\Gamma), \mathcal{J}^u_1)$–Patterson–Sullivan measure $\mu^{\mathcal{J}^u_1}$ on $\mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_2\}}(V_\mathbb{R})$ that projects to the Lebesgue measure on $\xi^1_\rho(\partial\Gamma)$. Since ρ is weakly irreducible, Lemma 6.5 implies that it is $\mu^{\mathcal{J}^u_1}$–irreducible, thus Theorem 5.14 applies to give

$$\mathsf{a}_1 = \mathcal{J}^u_1 \in \mathcal{D}_{\rho(\Gamma)}.$$

This is to say, $h_\rho(\mathsf{a}_1) \leq 1$. □

Before proceeding to arbitrary d_Γ we record a direct consequence of Corollary 6.6. Let us say that ρ is *coherent* if the first root arising in $\operatorname{span}(\xi^1_\rho(\partial\Gamma))$ is a_1.

Corollary 6.7 *Let Γ be a word-hyperbolic group such that $\partial\Gamma$ is homeomorphic to a circle. Let $\rho\colon \Gamma \to \mathsf{G}_\mathbb{K}$ be an a–Anosov representation and assume there exists a proximal, real representation $\Lambda\colon \mathsf{G}_\mathbb{K} \to \mathsf{PGL}(V_\mathbb{R})$ with first root a such that $\Lambda \circ \rho$ is coherent. Then*

$$\mathsf{a} \in \mathcal{Q}_{\rho(\Gamma)}.$$

6.4 When $\partial\Gamma$ has arbitrary dimension

Recall that a subgroup $\Gamma \subset \mathsf{PGL}(V_\mathbb{K})$ is *strongly irreducible* if any finite-index subgroup acts irreducibly. It is well known that this is equivalent to the fact that the connected component of the identity of the Zariski closure of Γ acts irreducibly on $\mathbb{K}^d$.

We will need the following lemma (which does not require Assumption 6.1):

Lemma 6.8 *Let $\eta\colon \Gamma \to \mathsf{PGL}_d(\mathbb{R})$ be a strongly irreducible a_1–Anosov representation, and assume that there exists $p \in [\![1, d-1]\!]$ and a measurable η–equivariant section $\zeta\colon \partial\Gamma \to \mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_p\}}(\mathbb{R}^d)$. Then η is μ^φ–irreducible for any $(\rho(\Gamma),\varphi)$–Patterson–Sullivan measure on $\mathcal{F}_\Theta(\mathbb{K}^d)$.*

Proof Otherwise, we would be able to find a subspace $W_0 \in \mathcal{F}_{\{\mathsf{a}_{d-p}\}}(\mathbb{R}^d)$ such that for almost every[7] $\xi^1_\rho(x) \in \xi^1_\rho(\partial\Gamma)$ one has $\zeta(x)^p \cap W_0 \neq \{0\}$. Since ζ is η–equivariant, we would find a p–dimensional subspace V such that, for every $\gamma \in \Gamma$,

$$\eta(\gamma)V \cap W_0 \neq \{0\}.$$

This implies that for every g in the Zariski closure of $\eta(\Gamma)$ it holds that $\dim gV \cap W_0 \geq 1$. The contradiction comes from Labourie [35, Proposition 10.3]: if G is an algebraic subgroup of $\mathsf{SL}(n,\mathbb{R})$, $C \in \mathcal{G}_k(\mathbb{R}^d)$, $B \in \mathcal{G}_{d-k}(\mathbb{R}^d)$ and $\dim(gC \cap B) \geq 1$ for every $g \in \mathsf{G}$, then the connected component of the identity of G is not irreducible. □

We can now prove Theorem A for arbitrary d_Γ.

Corollary 6.9 *Under Assumption 6.1 together with strong irreducibility of ρ,*

$$\mathcal{J}^u_{d_\Gamma} \in \mathcal{Q}_\rho(\Gamma).$$

Proof Since $\xi^1_\rho(\partial\Gamma)$ is a Lipschitz sphere it has Hausdorff dimension d_Γ, and thus Corollary 1.1 implies that $h_\rho(\mathcal{J}^u_{d_\Gamma}) \geq 1$. Proposition 6.4 guarantees the existence of a $(\rho(\Gamma),\mathcal{J}^u_{d_\Gamma})$–Patterson–Sullivan measure. Moreover, the equivariant map from (6-2) allows us to apply Lemma 6.8, and thus we have the hypothesis of Theorem 5.14. Consequently, $h_\rho(\mathcal{J}^u_{d_\Gamma}) \leq 1$. □

7 $(1,1,p)$–hyperconvex representations and a C^1–dichotomy for surface groups

In this section we will consider projective Anosov representations whose image of the boundary map is a C^1–submanifold. In the second part of the section we will prove Corollary 7.8, providing a C^1–dichotomy for surface groups.

[7]This is with respect to the pushed-forward measure $\pi_*\mu^\varphi$, where $\pi\colon \mathcal{F}_{\{\mathsf{a}_1,\mathsf{a}_p\}}(\mathbb{R}^d) \to \mathbb{P}(\mathbb{R}^d)$ consists of forgetting the p^{th} coordinate.

7.1 $(1,1,p)$–hyperconvex representations

Definition 7.1 We say a $\{\mathsf{a}_1,\mathsf{a}_p\}$–Anosov representation $\rho\colon\Gamma\to\mathsf{PGL}_d(\mathbb{R})$ is $(1,1,p)$–*hyperconvex* if, for every pairwise distinct $x,y,z\in\partial\Gamma$, the sum

$$\xi^1(x)+\xi^1(y)+\xi^{d-p}(z)$$

is direct.

Example 7.2 Zariski-dense hyperconvex representations can be obtained by deforming $\mathsf{S}^k\circ\iota$, where S^k denotes the k^{th} symmetric power and $\iota\colon\Gamma\to\mathsf{PO}(1,p)$ is the inclusion of a cocompact lattice; see Pozzetti, Sambarino and Wienhard [39, Corollary 7.6].

Hyperconvex representations were introduced by Labourie [35] for surface groups, and further studied by Zhang and Zimmer [48] when the boundary of Γ is topologically a sphere and by Pozzetti, Sambarino and Wienhard [39] for arbitrary hyperbolic groups. In both [39, Proposition 7.4] and [48, Theorem 1.1] one finds:

Theorem 7.3 *Assume that $\partial\Gamma$ is topologically a sphere of dimension $p-1$ and let $\rho\colon\Gamma\to\mathsf{PGL}_d(\mathbb{R})$ be a $(1,1,p)$–hyperconvex representation. Then $\xi^1_\rho(\partial\Gamma)$ is a C^1–sphere.*

Theorem A then gives:

Corollary 7.4 *Assume that $\partial\Gamma$ is topologically a sphere of dimension $p-1$ and let $\rho\colon\Gamma\to\mathsf{PSL}_d(\mathbb{R})$ be strongly irreducible and $(1,1,p)$–hyperconvex. Then $h_\rho(\mathfrak{J}^u_p)=1$.*

Remark 7.5 This generalizes Potrie and Sambarino [38, Corollary 7.1]. Observe however that, since the limit set $\xi^1(\partial\Gamma)$ is a C^1–submanifold of $\mathbb{P}(\mathbb{R}^d)$, the arguments of [38] adapt directly to give a version of Corollary 7.4 without requiring strong irreducibility.

Theorem 7.6 (Glorieux, Monclair and Tholozan [25]) *Let $\rho\colon\Gamma\to\mathsf{PGL}_d(\mathbb{R})$ be an a_1–Anosov representation that preserves a properly convex domain. Then*

$$2h_\rho(\omega_1+\omega_{d-1})\le\dim_{\mathrm{Hff}}((\xi^1,\xi^{d-1})(\partial\Gamma)),$$

where $(\xi^1,\xi^{d-1})\colon\partial\Gamma\to\mathbb{P}(\mathbb{R}^d)\times\mathbb{P}((\mathbb{R}^d)^*)$.

As an application of Corollary 7.4 we show that, for $(1,1,p)$–hyperconvex representations with $p<d-1$, such a bound can never be achieved:

Proposition 7.7 *Assume that $\partial\Gamma$ is topologically a sphere of dimension $p-1$ and let $\rho\colon\Gamma\to\mathsf{PGL}_d(\mathbb{R})$ be strongly irreducible and $(1,1,p)$–hyperconvex. If $p<d-1$, then*

$$2h_\rho(\omega_1+\omega_{d-1})<(1-\varepsilon)(p-1),$$

where $\varepsilon>0$ only depends on the $\{\mathsf{a}_1,\mathsf{a}_p\}$–Anosov constants of ρ.

Proof Since $p<d-1$ the functional $\phi\in\mathsf{E}^*$ given by

$$\phi=\frac{\omega_p-\omega_1}{p-1}-\frac{\omega_{d-1}-\omega_1}{d-2}$$

is nonzero. Moreover observe that, for every $v\in\mathsf{E}^+$,

$$\phi(v)\geq\frac{d-p-1}{d-2}\mathsf{a}_p(v).$$

Since ρ is a_p–Anosov, the last computation implies $\ker\phi\cap\mathcal{L}_{\rho(\Gamma)}=\{0\}$. This is to say that $\phi\in(\mathcal{L}_{\rho(\Gamma)})^*$, in particular ϕ has a well-defined entropy $h_\rho(\phi)\in(0,\infty)$. Then

$$h_\rho\biggl(\frac{p-1}{d-2}((d-1)\omega_1-\omega_{d-1})\biggr)=h_\rho(\mathfrak{J}^u_{p-1}+(p-1)\phi)\leq\frac{h_\rho(\phi)}{h_\rho(\phi)+p-1},\tag{7-1}$$

where the equality comes from the equality between the corresponding linear forms and the inequality follows from Lemma 5.6 together with Corollary 7.4 stating that $h_\rho(\mathfrak{J}^u_{p-1})=1$.

Finally, observe that

$$\begin{aligned}\tfrac{1}{2}(p-1)(\omega_1-\omega_{d-1})&=\frac{1}{2}\biggl(\frac{p-1}{d-2}((d-1)\omega_1-\omega_{d-1})+\frac{p-1}{d-2}((d-1)\omega_{d-1}-\omega_1)\biggr)\\&=\tfrac{1}{2}\bigl(\mathfrak{J}^u_{p-1}+(p-1)\phi+(\mathfrak{J}^u_{p-1}+(p-1)\phi)\circ\mathrm{i}\bigr),\end{aligned}$$

where $\mathrm{i}\colon\mathsf{E}\to\mathsf{E}$ is the opposition involution. Together with (7-1) and Lemma 5.6, this yields

$$\begin{aligned}\frac{2}{p-1}h_\rho(\omega_1-\omega_{d-1})&\leq2\frac{h_\rho(\mathfrak{J}^u_{p-1}+(p-1)\phi)h_\rho\bigl((\mathfrak{J}^u_{p-1}+(p-1)\phi)\circ\mathrm{i}\bigr)}{h_\rho(\mathfrak{J}^u_{p-1}+(p-1)\phi)+h_\rho\bigl((\mathfrak{J}^u_{p-1}+(p-1)\phi)\circ\mathrm{i}\bigr)}\\&=h_\rho(\mathfrak{J}^u_{p-1}+(p-1)\phi)\leq\frac{h_\rho(\phi)}{h_\rho(\phi)+p-1}<1,\end{aligned}$$

since entropy is i–invariant.

To conclude the proof we observe that the functional ϕ belongs to the Anosov–Levi space of every $\{\mathsf{a}_1,\mathsf{a}_p\}$–Anosov representation. Its entropy thus varies continuously

(Theorem 5.12), and hence
$$\eta \mapsto \frac{h_\eta(\phi)}{h_\eta(\phi)+p-1}$$
is bounded away from 1 on compact subsets of $\mathfrak{X}_{\{\mathsf{a}_1,\mathsf{a}_p\}}(\Gamma,\mathsf{PGL}_d(\mathbb{R}))$. $\square$

C^1–dichotomy

Now we prove the C^1–dichotomy announced in Section 1. As we will later see (Sections 9 and 10) there are many projective Anosov representations of surface groups where the image of the boundary map is Lipschitz. However, when we embed the surface group into $\mathsf{PSL}_2(\mathbb{R})$ and look at small deformations of representations
$$\Gamma \to \mathsf{PSL}_2(\mathbb{R}) \xrightarrow{R} \mathsf{PSL}_d(\mathbb{R}),$$
where R satisfies additional proximality assumptions ensuring that the representation is $\{\mathsf{a}_1,\mathsf{a}_2\}$–Anosov, then the image of the boundary map is never Lipschitz.

Recall that $g \in \mathsf{PGL}_d(\mathbb{R})$ is *proximal* if the generalized eigenspace associated to its greatest eigenvalue (in modulus) has dimension 1. A representation $R\colon G \to \mathsf{PGL}_d(\mathbb{R})$ of a given group G is *proximal* if its image contains a proximal element.

Corollary 7.8 *Let $R\colon \mathsf{PSL}_2(\mathbb{R}) \to \mathsf{PSL}_d(\mathbb{R})$ be a (possibly reducible) proximal representation such that $\bigwedge^2 R$ is also proximal. Let S be a closed connected surface of genus ≥ 2 and let $\rho_0\colon \pi_1 S \to \mathsf{PSL}_2(\mathbb{R})$ be discrete and faithful. Then we have the following dichotomy:*

(i) *If the top two weight spaces of R belong to the same irreducible factor, then for every small deformation $\rho\colon \pi_1 S \to \mathsf{PSL}_d(\mathbb{R})$ of $R\rho_0$ the curve $\xi^1_\rho(\partial\pi_1 S)$ is C^1.*

(ii) *Otherwise, for every weakly irreducible small deformation $\rho\colon \pi_1 S \to \mathsf{PSL}_d(\mathbb{R})$ of $R\rho_0$ the curve $\xi^1_\rho(\partial\pi_1 S)$ is not Lipschitz.*

Proof By the proximality assumptions on R, the representation
$$\rho := R\rho_0\colon \pi_1 S \to \mathsf{PSL}_d(\mathbb{R})$$
is $\{\mathsf{a}_1,\mathsf{a}_2\}$–Anosov: indeed, $\mathsf{PSL}(2,\mathbb{R})$ has rank one. This implies, on the one hand, that the discrete and faithful representation ρ_0 is Anosov, and on the other hand that the composition of ρ_0 with any proximal representation is a_1 Anosov.

Furthermore, if the first two weights of R belong to the same irreducible factor, the representation ρ is also $(1,1,2)$–hyperconvex [39, Proposition 6.16]. Hyperconvexity is an open property in $\mathfrak{X}(\pi_1 S, \mathsf{PSL}_d(\mathbb{R}))$ (Pozzetti, Sambarino and Wienhard [39]) and thus Theorem 7.3 implies that every small deformation of ρ has C^1 limit set.

If instead the two top weights of R belong to different irreducible factors, then it follows from the representation theory of $\mathsf{SL}_2(\mathbb{R})$ that

$$h_\rho(\mathsf{a}_1) = h_\rho(\mathfrak{J}_1^u) = 2.$$

Note that the entropy of $\mathfrak{J}_1^u$ is continuous on $\mathfrak{X}_{\{\mathsf{a}_1,\mathsf{a}_2\}}(\pi_1 S, \mathsf{PSL}_d(\mathbb{R}))$; see Theorem 5.12. In particular there exists a neighborhood $\mathcal{U}$ of ρ such that $h_\eta(\mathfrak{J}_1^u) > 1$ for every $\eta \in \mathcal{U}$. Theorem A implies that no weakly irreducible representation in $\mathcal{U}$ can have Lipschitz limit set. $\square$

The regular case, Corollary 7.8(i), is inspired by Labourie [35], who treated the case (of arbitrary deformations) of the irreducible representations, and was proven in Pozzetti, Sambarino and Wienhard [39, Proposition 9.4]. The novelty of this paper is item (ii), inspired by Barbot [1], who proved it for $d = 3$. We believe both items placed together give a clearer picture.

It is easy to obtain similar results for other groups G by considering suitable linear representations. On the other hand, the double proximality assumption is necessary: the composition of a maximal representation not in the Hitchin component and the irreducible linear representation of $\mathsf{Sp}(2n, \mathbb{R})$ of highest weight w_n is proximal, but its second exterior power is not proximal. It is possible to check that no small Zariski-dense deformation satisfies either (i) or (ii).

Along the same lines we can deduce that some natural Anosov representations of hyperbolic lattices do not have Lipschitz boundary maps:

Corollary 7.9 *Let $\Gamma < \mathsf{PO}(1,n)$ be a lattice, $n \geq 3$ and $\rho_1 \colon \Gamma \to \mathsf{PO}(1,m)$ strictly dominated by the lattice embedding ρ_0. Then, for any Zariski-dense small deformation of $\rho_0 \oplus \rho_1^{n-1}$, the limit set $\xi_\rho^1(\partial\Gamma)$ is not Lipschitz.*

Examples of lattices Γ admitting such representations were constructed by Danciger, Guéritaud and Kassel [17, Proposition 1.8].

8 $\mathbb{H}^{p,q}$ convex–cocompact representations

Generalizing work of Mess [37] and Barbot and Mérigot [3], Danciger, Guéritaud and Kassel [16] introduced a class of representations called $\mathbb{H}^{p,q}$*–convex–cocompact*.

These form another interesting class of representations with Lipschitz boundary map where Theorem A applies.

Let $d = p + q$ with $p, q \geq 1$ and let Q be a symmetric bilinear form on $\mathbb{R}^d$ of signature (p, q). The subspace of $\mathbb{P}(\mathbb{R}^d)$ consisting of negative definite lines is called the *pseudo-Riemannian hyperbolic space* and denoted by

$$\mathbb{H}^{p,q-1} = \{\ell \in \mathbb{P}(\mathbb{R}^d) : Q|_{\ell\setminus\{0\}} < 0\}.$$

The cone of isotropic lines is usually denoted by $\partial\mathbb{H}^{p,q-1}$.

Instead of the original definition of convex–cocompactness, we recall the characterization given by [16, Theorem 1.11]:

Definition 8.1 An a_1–Anosov representation $\rho\colon \Gamma \to \mathsf{PO}(p, q)$ is $\mathbb{H}^{p,q-1}$*–convex–cocompact* if, for every pairwise distinct triple of points $x, y, z \in \partial\Gamma$, the restriction $Q|_{\xi^1_\rho(x)\oplus\xi^1_\rho(y)\oplus\xi^1_\rho(z)}$ has signature $(2, 1)$.

When Γ_0 is a cocompact lattice in $\mathsf{SO}(p, 1)$, $\mathbb{H}^{p,1}$–convex–cocompact representations of Γ_0 are usually referred to as AdS–*quasi-Fuchsian groups*. Barbot [2] proved that these groups form connected components of the character variety $\mathfrak{X}(\Gamma_0, \mathsf{SO}(p, 2))$ only consisting of Anosov representations. In [23] Glorieux and Monclair prove that the limit set of an AdS–quasi-Fuchsian group is never a C^1–submanifold, except for Fuchsian groups.

The following is well known and easy to verify; see for example Glorieux and Monclair [24, Proposition 5.2].

Proposition 8.2 *Assume that $\partial\Gamma$ is homeomorphic to a $(p-1)$–dimensional sphere. If $\rho\colon \Gamma \to \mathsf{PO}(p, q)$ is $\mathbb{H}^{p,q}$–convex–cocompact, then $\xi^1_\rho(\partial\Gamma)$ is a Lipschitz submanifold of $\partial\mathbb{H}^{p,q-1}$.*

Proof The space $\partial\mathbb{H}^{p,q-1}$ admits a twofold cover that splits as $\mathbb{S}^{p-1} \times \mathbb{S}^{q-1}$. It is immediate to verify that, since for every pairwise distinct triple $(x, y, z) \in \partial\Gamma$ we have that $Q|_{\xi^1_\rho(x)\oplus\xi^1_\rho(y)\oplus\xi^1_\rho(z)}$ has signature $(2, 1)$, each one of the two lifts of $\xi^1_\rho(\partial\Gamma)$ to $\mathbb{S}^{p-1} \times \mathbb{S}^{q-1}$ is the graph of a 1–Lipschitz function $f : \mathbb{S}^{p-1} \to \mathbb{S}^{q-1}$, and, as such, is a Lipschitz submanifold of $\partial\mathbb{H}^{p,q-1}$. □

Theorem A then yields:

Corollary 8.3 *Assume that $\partial\Gamma$ is homeomorphic to a $(p-1)$–dimensional sphere and let $\rho\colon \Gamma \to \mathsf{PO}(p,q)$ be $\mathbb{H}^{p,q-1}$–convex–cocompact. Then*

- *for $p = 2$ and ρ weakly irreducible, $h_\rho(\mathfrak{J}_1^u) = 1$;*
- *for $p \geq 3$ and ρ strongly irreducible, $h_\rho(\mathfrak{J}_{p-1}^u) = 1$.*

One concludes the following upper bound for the entropy of the spectral radius inspired by Glorieux and Monclair [24].

Corollary 8.4 *Assume that $\partial\Gamma$ is homeomorphic to a $(p-1)$–dimensional sphere and let $\rho\colon \Gamma \to \mathsf{PO}(p,q)$ be $\mathbb{H}^{p,q-1}$–convex–cocompact. Then*

- *for $p = 2$ and ρ weakly irreducible, $h_\rho(\omega_1) \leq 1$;*
- *for $p \geq 3$ and ρ strongly irreducible, $h_\rho(\omega_1) \leq p-1$.*

Proof Assume first that $p \leq q$ and note that, for every $g \in \mathsf{PO}(p,q)$,

$$(\omega_p - \omega_1)(\lambda(g)) = \lambda_2(g) + \cdots + \lambda_p(g) \geq 0.$$

By definition, $\mathfrak{J}_{p-1}^u = p\omega_1 - \omega_p$, and thus

$$\frac{h_\rho(\omega_1)}{p-1} = h_\rho((p-1)\omega_1) \leq h_\rho(\mathfrak{J}_{p-1}^u) = 1$$

by Corollary 8.3. The only difference in the case where $q < p$ is that $\mathfrak{J}_{p-1}^u = p\omega_1 - \omega_q$, but the same argument applies verbatim. □

The entropy for the first fundamental weight has a particular meaning for projective Anosov representations into $\mathsf{PO}(p,q)$, notably for $q \geq 2$. Fix $o \in \mathbb{H}^{p,q-1}$ and consider

$$S^o = \{W < \mathbb{R}^d : o \subset W, \dim W = q \text{ and } Q|_W \text{ is negative definite}\}.$$

This is a totally geodesic embedding of the symmetric space $X_{p,q-1}$ of $\mathsf{PO}(p,q-1)$ in the symmetric space $X_{p,q}$.

Given a projective Anosov representation $\rho\colon \Gamma \to \mathsf{PO}(p,q)$ one defines the open subset of $\mathbb{H}^{p,q-1}$

$$\Omega_\rho = \{o \in \mathbb{H}^{p,q-1} : Q(o, \xi_\rho^1(x)) \neq 0 \text{ for all } x \in \partial\Gamma\}.$$

Carvajales [13] shows that, assuming $\Omega_\rho \neq \varnothing$, for every $o \in \Omega_\rho$ one has

$$\lim_{t\to\infty} \frac{\log \#\{\gamma \in \Gamma : d_{X_{p,q}}(S^o, \rho(\gamma)S^o)\}}{t} = h_\rho(\omega_1),$$

and provides an asymptotic for this counting function; see [13, Theorem A].

When ρ is also $\mathbb{H}^{p,q-1}$–convex–cocompact, Glorieux and Monclair [24, Section 1.2] introduce a *pseudo-Riemannian critical exponent* δ_ρ, and show, in particular, that

$$\delta_\rho \leq p-1$$

[24, Theorem 1.2]. Carvajales proves [13, Remarks 6.9 and 7.15] that $\delta_\rho = h_\rho(\omega_1)$, so Corollary 8.4 provides a different proof of [24, Theorem 1.2] when Γ is assumed to have boundary homeomorphic to a $(p-1)$–dimensional sphere.

We finish the section with a direct application of Theorem 5.4 and Corollary 8.3 allowing us to get a bound for the Riemannian critical exponent. We use freely the notation from Remark 5.5.

Consider a representation $\Lambda\colon \mathsf{PO}(p,1) \to \mathsf{PO}(p,q)$ such that its image stabilizes a $(p+1)$–dimensional subspace V of $\mathbb{R}^d$ where $Q|_V$ has signature $(p,1)$. Endow the symmetric space $X_{p,q}$ with a $\mathsf{PO}(p,q)$–invariant Riemannian metric such that the totally geodesic copy of $\mathbb{H}^p$ in $X_{p,q}$ induced by Λ has constant curvature -1. In particular, if $\iota\colon \Gamma \to \mathsf{PO}(p,1)$ is the lattice embedding, $h^X_{\Lambda\circ\iota} = p-1$. We show that this is an upper bound for any strongly irreducible, $\mathbb{H}^{p,q-1}$–convex–cocompact representation:

Proposition 8.5 *Assume that $\partial\Gamma$ is homeomorphic to a $(p-1)$–dimensional sphere, and let $\rho\colon \Gamma \to \mathsf{PO}(p,q)$ be strongly irreducible and $\mathbb{H}^{p,q-1}$–convex–cocompact. Then*

$$h^X_\rho \leq p-1.$$

Proof In view of Theorem 5.4 (or more precisely Remark 5.5), it suffices to recall that $\mathcal{D}_{\rho(\Gamma)}$ is convex (Proposition 5.3) and that, by Corollary 8.3,

$$\mathcal{J}^u_{p-1} \in \mathcal{Q}_{\rho(\Gamma)}.$$

See Potrie and Sambarino [38, Section 1.1] for more details. □

9 Maximal representations

An important class of representations that are in general only Anosov with respect to one maximal parabolic subgroup but admit boundary maps with Lipschitz image are maximal representations into Hermitian Lie groups. In this case the Lipschitz property for the image of the boundary map is a consequence of a positivity/causality property of the boundary map. We first describe the causal structure on the Shilov boundary of a Hermitian symmetric space of tube type, introduce the notion of a positive curve and show that the image of any positive curve (that is not necessarily equivariant with respect to a representation) is a Lipschitz submanifold. We then show how this applies to maximal representations and allows us to prove Theorem 9.8, the main result of this section. We also deduce consequences for the orbit growth rate on the symmetric space.

9.1 Causal structure and positive curves

Let $\mathsf{G}_{\mathbb{R}}$ be a simple Hermitian Lie group of tube type. Examples to keep in mind are the symplectic group $\mathsf{G}_{\mathbb{R}} = \mathsf{Sp}(2n, \mathbb{R})$ or the orthogonal group $\mathsf{G}_{\mathbb{R}} = \mathsf{SO}_0(2, n)$. The Shilov boundary $\check{S}$ of the bounded domain realization of the symmetric space associated to $\mathsf{G}_{\mathbb{R}}$ is a flag variety $\mathsf{G}_{\mathbb{R}}/\check{P}$, where $\check{P}$ is a maximal parabolic subgroup determined by a specific simple root $\{\check{\mathsf{a}}\}$. In the first of our main examples, $\mathsf{G}_{\mathbb{R}} = \mathsf{Sp}(2n, \mathbb{R})$, the parabolic subgroup $\check{P}$ in question is the stabilizer of a Lagrangian subspace $L \in \mathscr{L}(\mathbb{R}^{2n})$ so $\check{\mathsf{a}} = \mathsf{a}_n$, and in the second, $\mathsf{G}_{\mathbb{R}} = \mathsf{SO}_0(2, n)$, $\check{P}$ is the stabilizer of an isotropic line $l \in \mathsf{Is}_1(\mathbb{R}^{2,n})$, so $\check{\mathsf{a}} = \mathsf{a}_1$.

In general, for a simple Hermitian Lie group of rank n, there is a special set of n *strongly orthogonal roots* $\mathsf{b}_1, \dots, \mathsf{b}_n$ of the complexification $\mathfrak{g}_{\mathbb{C}}$; see [29, pages 582–583]. The set of strongly orthogonal roots give rise to a (holomorphic) embedding of a maximal polydisk. If the symmetric space is of tube type, the simple root $\check{\mathsf{a}}$ is the smallest strongly orthogonal root $\check{\mathsf{a}} = \mathsf{b}_n$. All the other strongly orthogonal roots are of the form $\mathsf{b}_i = \mathsf{b}_n + \varphi$, where $\varphi \in \mathsf{E}^*$ is nonnegative on the Weyl-chamber. We record the following for later use:

Lemma 9.1 *Let $a \in \mathsf{E}^+$. Then $\check{\mathsf{a}}(a) = \min_{i=1,\dots,n} \mathsf{b}_i(a)$.*

For Hermitian groups of tube type, the Shilov boundary carries a natural causal structure: for every $p \in \check{S}$ there is an open convex acute cone $C_p \subset \mathsf{T}_p\check{S}$, which we now define.

Recall that $\mathsf{G}_{\mathbb{R}}/\check{P}$ can be identified as the space of parabolic subgroups of $\mathsf{G}_{\mathbb{R}}$ that are conjugate to $\check{P}$. Let us fix a point $\check{p} = \check{P} \in \check{S}$, which one should think of as a point at infinity. Then, at any point $p = P \in \check{S}$ that is transverse to $\check{p}$, ie such that the parabolic groups P and $\check{P}$ are opposite, the tangent space $\mathsf{T}_p\check{S}$ is identified with the Lie algebra $\check{\mathfrak{n}}$ of the unipotent radical of $\check{P}$, and the cone C_p is an open convex acute cone $\check{C} \subset \check{\mathfrak{n}}$, invariant under the action of the connected component of $P \cap \check{P}$.

In the case of $\mathsf{Sp}(2n, \mathbb{R})$ this is the cone of positive definite symmetric matrices, and in the case of $\mathsf{SO}_0(2, n)$ it is the cone of vectors with positive first entry that are positive for the induced conformal class of Lorentzian inner products on $\mathsf{T}_P\,\mathsf{Is}_1(\mathbb{R}^{2,n})$.

This invariant cone $\check{C} \subset \check{\mathfrak{n}}$ in fact also gives rise to the notion of maximal triples in $\check{S}$ via the exponential map. A triple $(P, Q, \check{P})$ is said to be maximal if there exists an $s \in \check{C}$ such that $Q = \exp s \cdot P$. Extending this by the action of G leads to a notion of maximal triples in $\check{S}$, which actually coincides exactly with those triples which have maximal (generalized) Maslov index as introduced by Clerc and Ørsted [14].

Definition 9.2 Let $\check{S}$ be the Shilov boundary of a Hermitian symmetric space of tube type. A curve $\xi\colon \mathbb{S}^1 \to \check{S}$ is *positive* if the image of any positively oriented triple is a maximal triple.

Proposition 9.3 *Let $\xi\colon \mathbb{S}^1 \to \check{S}$ be a positive curve. Then $\xi(\mathbb{S}^1)$ is a Lipschitz submanifold of $\check{S}$.*

Proof Note that whenever we pick two points $p_1 = P_1$ and $p_2 = P_2$ on the image of ξ, the image $\xi(\mathbb{S}^1)$ can be covered by the two charts consisting of parabolic subgroups that are transverse to p_1 and p_2, respectively.

In any of these charts the inverse image of ξ, under the exponential map

$$\mathfrak{n}_i \to \mathsf{G}_{\mathbb{R}}/P_i, \quad s \mapsto \exp(s)\check{P}_j,$$

gives a map $\bar{\xi}\colon \mathbb{R} \to \mathfrak{n}_i$ such that, for every $t_1 < t_2$, we have that $\bar{\xi}(t_2) - \bar{\xi}(t_1)$ is contained in the open convex acute cone $\check{C}$. It then follows (see for example Burger, Iozzi, Labourie and Wienhard [10, Lemma 8.10]) that the restriction of $\bar{\xi}$ to any bounded interval has finite length. As a result, $\xi(\mathbb{S}^1) \subset \check{S}$ is rectifiable. It is thus possible to reparametrize $\mathbb{S}^1$ so that ξ is a Lipschitz map. □

Remark 9.4 We did not assume that the positive map is equivariant with respect to a representation. This will be important in Section 10, where we will apply Proposition 9.3 in this generality.

9.2 Maximal representations

Let G denote a Hermitian semisimple Lie group and let Γ denote the fundamental group of a closed hyperbolic surface S. We consider representations $\rho\colon \Gamma \to \mathsf{G}$ that are maximal, in the sense that they maximize the Toledo invariant, whose definition was recalled in Section 1. Important for us is that they can be characterized in terms of boundary maps:

Theorem 9.5 (Burger, Iozzi and Wienhard [12, Theorem 8]) *A representation $\rho\colon \Gamma \to \mathsf{G}$ is maximal if and only if there exists a continuous, ρ–equivariant, positive map $\phi\colon \partial\Gamma \to \check{S}$.*

In order to apply Corollary 6.7 we need to verify some weak irreducibility assumptions. Let us first treat the case when the Zariski closure of $\rho(\Gamma)$ is simple.

Corollary 9.6 *Let* G *be a simple Hermitian Lie group of tube type and let* $\check{\mathsf{a}}$ *be the root associated the Shilov boundary of* G*. If* $\rho\colon\Gamma\to\mathsf{G}$ *is a Zariski-dense maximal representation, then*

$$\check{\mathsf{a}}\in\mathcal{Q}_{\rho(\Gamma)},$$

or equivalently $h_\rho(\check{\mathsf{a}})=1$.

Proof The proof follows from Corollary 6.7 and Proposition 9.3 by considering the representation $\Lambda_{\check{a}}$ from Proposition 4.4. □

In the remainder of this section we show how the case of maximal representations with semisimple target group that are not necessarily Zariski-dense can be reduced to Corollary 9.6. To this end we will use a result from Burger, Iozzi and Wienhard [11] describing the Zariski closure H of a maximal representation: H splits as $\mathsf{H}_1\times\cdots\times\mathsf{H}_n$, each factor is Hermitian, and the inclusion in $\mathsf{H}\to\mathsf{G}$ is *tight*. In the following we will not need the definition of a tight homomorphism, and therefore refer the interested reader to [11, Definition 1].

The following lemma will then be useful:

Lemma 9.7 *Let* G *be a classical simple Hermitian Lie group of tube type and consider a tight embedding* $\iota\colon\mathsf{H}=\mathsf{H}_1\times\cdots\times\mathsf{H}_n\to\mathsf{G}$*. If we denote by* $\iota_*\colon\mathsf{E}^+_{\mathsf{H}}\to\mathsf{E}^+_{\mathsf{G}}$ *the induced map, then*

$$\check{\mathsf{a}}_{\mathsf{G}}\circ\iota_*=\min_i\check{\mathsf{a}}_{\mathsf{H}_i}.$$

Proof Denote by $\pi\colon\mathfrak{h}_1\oplus\cdots\oplus\mathfrak{h}_n\to\mathfrak{g}$ the associated Lie algebra homomorphism. Let E_i be a Cartan subspace of H_i and E_{G} a Cartan subspace of G such that $\pi(\mathsf{E}_i)\subset\mathsf{E}_{\mathsf{G}}$.

As ι is tight and G is classical, the classification of Hamlet and Pozzetti [28] applies and gives that we have an orthogonal decomposition $\mathsf{E}_{\mathsf{G}}=\mathsf{B}_1\oplus\cdots\oplus\mathsf{B}_k$ such that $\pi|_{\bigoplus\mathsf{E}_i}$ is a direct sum of maps $\pi_i\colon\mathsf{E}_i\to\mathsf{B}_i$. Furthermore, there are only a few possibilities for the linear map π_i. If H_i has rank greater than one, then $\mathsf{B}_i=\mathsf{E}_i^{m_i}$ for some m_i and π_i is a diagonal inclusion; if instead E_i is one-dimensional, or equivalently $\mathsf{H}_i\cong\mathsf{PSL}_2(\mathbb{R})$, then π_i is induced from a direct sum of nontrivial irreducible representations (of varying degrees). It is easy to check that the subspace B_i is then the span of the real vectors in $\mathfrak{p}$ associated to the strongly orthogonal roots that do not vanish on $\pi(\mathsf{E}_i)$. Setting $\mathsf{b}_i=\min_{j,\mathsf{b}_j|_{\mathsf{E}_i}\neq0}\mathsf{b}_j$, we have $\mathsf{b}_i|_{\pi(\mathsf{E}_i)}=\check{\mathsf{a}}_{\mathsf{H}_i}$. And hence, with Lemma 9.1, we have $\check{\mathsf{a}}_{\mathsf{G}}=\min_i(\check{\mathsf{a}}_{\mathsf{H}_i})$. □

Theorem 9.8 *Let G be a Hermitian semisimple Lie group such that all factors of G that are of tube type are classical. Let $\theta \subset \Delta$ be the subset of simple roots associated to the Shilov boundary of G. Then for every maximal representation $\rho\colon \Gamma \to \mathsf{G}$,*

$$\theta \subset \mathcal{Q}_{\rho(\Gamma)}.$$

Proof If $\mathsf{G} = \mathsf{G}_1 \times \cdots \times \mathsf{G}_n$ then $\check{S} = \check{S}_1 \times \cdots \times \check{S}_n$, and so $\theta = \{\check{\mathsf{a}}_{\mathsf{G}_1}, \dots, \check{\mathsf{a}}_{\mathsf{G}_n}\}$; see Burger, Iozzi and Wienhard [11, Lemma 3.2(1)]. Furthermore $\rho\colon \Gamma \to \mathsf{G}$ is maximal if and only if all $\rho_i : \Gamma \to \mathsf{G}_i$ are maximal (see Burger, Iozzi and Wienhard [12, Lemma 6.1(3)]). Therefore we can restrict to the case that G is simple.

Since every maximal representation factors through a representation into the normalizer of a maximal tube type subgroup $\mathsf{H} < \mathsf{G}$ (Burger, Iozzi and Wienhard [12, Theorem 5(3)]), which is simple, has the same rank as G, and is such that $\check{\mathsf{a}}_{\mathsf{G}} = \check{\mathsf{a}}_{\mathsf{H}}$, we can restrict to the tube type case as the limit set in $\check{S}_{\mathsf{G}}$ is contained in $\check{S}_{\mathsf{H}}$ and coincides with the limit set in $\check{S}_{\mathsf{H}}$. The maximal tube type domains are always classical Hermitian symmetric spaces, except for the one exceptional Hermitian symmetric space of tube type.

If now ρ is not Zariski-dense, then the Zariski closure is reductive and of tube type, so it is of the form $\mathsf{H}_1 \times \cdots \times \mathsf{H}_n$ and the representations into H_i are Zariski-dense and maximal. Therefore we have $h_\rho(\check{\mathsf{a}}_{\mathsf{H}_i}) = 1$ for all i. As the inclusion $\mathsf{H}_1 \times \cdots \times \mathsf{H}_n \to \mathsf{G}$ is tight, the result follows from Lemmas 9.7 and 5.1. □

9.3 Application to the Riemannian critical exponent

Any simple Hermitian Lie group G admits a diagonal embedding $\iota^\Delta\colon \mathsf{SL}_2(\mathbb{R}) \to \mathsf{G}$, which is equivariant with the inclusion of a diagonal disk in a maximal polydisk. We say that a representation $\rho\colon \Gamma \to \mathsf{G}$ is *diagonal-Fuchsian* if it has the form $\rho = \iota^\Delta \circ \rho_0$, where $\rho_0\colon \Gamma \to \mathsf{SL}_2(\mathbb{R})$ is the lift of the holonomy of a hyperbolization.

Let $K_\Delta < \mathsf{G}$ be the centralizer of the image of ι^Δ, which is compact. Then a diagonal Fuchsian representation ρ can be twisted by a representation $\chi\colon \Gamma \to K_\Delta$. We call the corresponding representation $\rho_\chi : \Gamma \to \mathsf{G}$ a *twisted diagonal* representation. Observe that the Riemannian critical exponent h^X is constant on twisted diagonal representations (the exact value h^X_{diag} depends on the choice of the normalization of the Riemannian metric)

Proposition 9.9 *Let Γ be the fundamental group of a closed surface and let $\rho\colon \Gamma \to \mathsf{G}$ be a maximal representation. Then $h^X_\rho \le h^X_{\mathrm{diag}}$.*

Proof Let $\mathsf{b}_1, \dots, \mathsf{b}_n$ be the set of strongly orthogonal roots for $\mathsf{G}_{\mathbb{C}}$. It is immediate to verify that the limit cone $\mathcal{L}_{\rho_0(\Gamma)}$ of a representation ρ_0 in the Fuchsian locus is concentrated in the span of the vertex of the Weyl chamber $\sum_{i=1}^n \mathsf{b}_i^*$, where b^* is the basis of E dual to $\{\mathsf{b}_1, \dots, \mathsf{b}_n\}$. We know from Corollary 9.6 that, for every ρ, the growth rate $h_\rho(\check{\mathsf{a}})$ equals 1. Thus, if we denote by $(\mathsf{E}^+)^*$ the cone of functionals that are nonnegative on the Weyl chamber, we get that $\check{\mathsf{a}} + (\mathsf{E}^+)^* \subset \mathcal{D}_{\rho(\Gamma)}$, and in particular all the strongly orthogonal roots are in $\mathcal{D}_{\rho(\Gamma)}$. A simple computation shows that the affine simplex determined by the strongly orthogonal roots meets the ray $\mathbb{R} \sum_{i=1}^n \mathsf{b}_i$ orthogonally in a point (it is just the diagonal in a positive quadrant meeting the span of the basis vectors), whose norm has to compute the Riemannian orbit growth rate of any representation ρ_0 in the Fuchsian locus: $\mathcal{Q}_{\rho_0(\Gamma)}$ is the affine hyperplane orthogonal to $\mathbb{R} \sum_{i=1}^n \mathsf{b}_i$ that contains $\check{\mathsf{a}}$. Remark 5.5 concludes the proof. □

Remark 9.10 When G is $\mathsf{Sp}(4, \mathbb{R})$, or more generally $\mathsf{SO}_\circ(2, n)$, it follows from Collier, Tholozan and Toulisse [15] that the bound is furthermore rigid: the equality is strict unless ρ is equal to ρ_0 up to a character in the compact centralizer of its image.

Note that for maximal representations into $\mathsf{Sp}(2n, \mathbb{R})$, for $n \geq 3$, every connected component of the space of maximal representations contains a twisted diagonal representation. However for $\mathsf{Sp}(4, \mathbb{R})$ there are exceptional components, discovered by Gothen, where every representation is Zariski-dense; see Bradlow, García-Prada and Gothen [8] and Guichard and Wienhard [26]. In these components it is easy to verify that the bound we provide is sharp, despite not being achieved.

In the special case of the Hitchin component of $\mathsf{Sp}(2n, \mathbb{R})$ the bound of Proposition 9.9 is never attained, as the irreducible representations provide a better bound that is furthermore rigid; see Potrie and Sambarino [38].

10 Θ–positive representations

Throughout this section we will write

$$\mathsf{G} = \mathsf{SO}(p, q),$$

with $p < q$. We consider the subset $\Theta = \{\mathsf{a}_1, \dots, \mathsf{a}_{p-1}\}$ of the simple roots discussed in Example 4.6 and denote by P_Θ the corresponding parabolic group, by L_Θ its Levi factor and by U_Θ its unipotent radical.

The group G admits a Θ*–positive structure* as defined by Guichard and Wienhard [27]. This means that for every $\mathsf{b} \in \Theta$ there exists an L^0_Θ–invariant sharp convex cone c_b in

$$\mathfrak{u}_\mathsf{b} = \sum_{\substack{\mathsf{a} \in \Sigma^+_\Theta \\ \mathsf{a}=\mathsf{b} \bmod \mathrm{Span}(\Pi \setminus \Theta)}} \mathfrak{g}_\mathsf{a}.$$

Here $\Sigma^+_\Theta = \Sigma^+ \setminus \mathrm{Span}(\Pi \setminus \Theta)$. For $\mathsf{b} \in \{\mathsf{a}_1, \dots, \mathsf{a}_{p-2}\}$, the space $\mathfrak{u}_\mathsf{b}$ is one-dimensional and the sharp convex cone $c_\mathsf{b} = \mathbb{R}^+ \subset \mathbb{R}$ consists of the positive elements, while $\mathfrak{u}_{\mathsf{a}_{p-1}} = \mathbb{R}^{q-p+2}$ is endowed with a form q_J of signature $(1, q-p+1)$ preserved by the action of $\mathsf{L}^0_\Theta = \mathbb{R}^{p-2} \times \mathsf{SO}^0(1, q-p+1)$. The cone $c_{\mathsf{a}_{p-1}}$ consists precisely of the positive vectors for q_J whose first entry is positive.

Following [27, Section 4.3] we denote by $W(\Theta)$ the subgroup of the Weyl group W generated by the reflections $\{\sigma_i\}_{i=1}^{p-2}$ together with the longest element σ_{p-1} of the Weyl group $W_{\mathsf{a}_{p-1},\mathsf{a}_p}$ of the subroot system generated by the last two simple roots. $W(\Theta)$ is, in our case, a Weyl group of type B_{p-1}. We denote by w^0_Θ the longest element of $W(\Theta)$, and choose a reduced expression $w^0_\Theta = \sigma_{i_1} \cdots \sigma_{i_l}$. Of course every reflection σ_i appears at least once among the σ_{i_k}. We consider the map

$$F_{\sigma_{i_1} \cdots \sigma_{i_l}} \colon c^0_{\mathsf{a}_{i_1}} \times \cdots \times c^0_{\mathsf{a}_{i_l}} \to U_\Theta, \quad (v_1, \dots, v_l) \mapsto \exp(v_1) \cdots \exp(v_l).$$

The Θ–positive semigroup U^+_Θ is defined as the image of the map $F_{\sigma_{i_1} \cdots \sigma_{i_l}}$, and doesn't depend on the choice of the reduced expression [27, Theorem 4.5].

A θ–positive structure on G gives rise to the notion of a positive triple in G/P_Θ.

Definition 10.1 A pairwise transverse triple in $(\mathsf{G}/P_\Theta)^3$ is Θ*–positive* if it lies in the G–orbit of a triple of the form $(F_1, u \cdot F_1, F_3)$, where $\mathrm{Stab}(F_3) = P_\Theta$, F_1 is transverse to F_3 and $u \in U^+_\Theta$ [27, Definition 4.6].

Remark 10.2 The stabilizer in $\mathsf{SO}^0(1, q-p+1)$ of a vector $v \in c_{\mathsf{a}_{p-1}}$ is compact. As a result one readily checks that the stabilizer in G of a Θ–positive triple is compact.

Let Γ_g be the fundamental group of a hyperbolic surface. A representation $\rho\colon \Gamma_g \to \mathsf{G}$ is Θ*–positive* if there exists a ρ–equivariant map $\partial\Gamma_g \to \mathsf{G}/P_\Theta$ sending positive triples to Θ–positive triples [27, Definition 5.3]. Guichard, Labourie and Wienhard show that every Θ–positive representation is necessarily Θ–Anosov [27, Conjecture 5.4], but since the proof has not yet appeared in print, in this section we will freely add this last assumption, and only discuss Θ–positive Anosov representations.

Theorem 10.3 *Let $\rho\colon \Gamma_g \to \mathsf{SO}(p,q)$ be Θ–positive and Θ–Anosov. For every $1 \le k \le p-2$ the representation $\bigwedge^k \rho$ is $(1,1,2)$–hyperconvex.*

Proof We denote by $\xi\colon \partial\Gamma_g \to \mathsf{G}/P_\Theta$ the Θ–positive continuous equivariant boundary map, and by $\xi^i : \partial\Gamma_g \to \mathsf{Is}_i(\mathbb{R}^{p,q})$ the induced maps. Let $(x,y,z) \in \partial^3\Gamma$ be a positively oriented triple. By assumption, $\xi(y) = s \cdot \xi(x)$ for some element s in the positive semigroup of the unipotent radical of the stabilizer of $\xi(z)$. In turn, $s = \exp(v_1)\cdots\exp(v_l)$ with $v_t \in c^0_{\mathsf{a}_{i_t}}$ (recall that $i_t \in \{1,\dots,p-1\}$).

We set $d = p+q$. It follows from [39, Proposition 8.11] that, in order to check that $\bigwedge^k \rho$ is $(1,1,2)$–hyperconvex, it is enough to verify that the sum

$$\xi^k_\rho(x) + (\xi^k_\rho(y) \cap \xi^{d-k+1}_\rho(z)) + \xi^{d-k-1}_\rho(z)$$

is direct, or equivalently that the sum

$$\xi^k_\rho(x) + s \cdot (\xi^k_\rho(x) \cap \xi^{d-k+1}_\rho(z)) + \xi^{d-k-1}_\rho(z)$$

is direct (recall that s belongs to the stabilizer of $\xi_\rho(z)$). Without loss of generality we can assume that the form Q defining the group $\mathsf{SO}(p,q)$ is represented by

$$Q = \begin{pmatrix} 0 & 0 & K \\ 0 & J & 0 \\ K^t & 0 & 0 \end{pmatrix},$$

with

$$K = \begin{pmatrix} 0 & 0 & (-1)^{p-1} \\ 0 & \iddots & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -\mathrm{Id}_{q-p} & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

We can furthermore assume that $\xi^l(z) = \langle e_1,\dots,e_l\rangle$ and $\xi^l(x) = \langle e_d,\dots,e_{d-l+1}\rangle$, so that $\xi^k(x) \cap \xi^{d-k+1}(z) = e_{d-k+1}$. In order to check that the representation is $(1,1,2)$–hyperconvex, we only have to verify that, given s as above, writing $s \cdot e_{d-k+1} = \sum \alpha_i e_i$, the coefficient α_{d-k} never vanishes. We claim that such coefficient is just $\sum_{i_t=k} v_t > 0$. Indeed, by construction, if $v_t \in c^0_{\mathsf{a}_m}$ with $m \in \{1,\dots,p-2\}$, then $\exp(v_t) \in \mathsf{SO}(p,q)$ differs from the identity only in the positions $(t,t+1)$ and $(d-t,d-t+1)$ where it is equal to v_t (see [27, Section 4.5]), while if $v_t \in c^0_{\mathsf{a}_{p-1}}$,

$$\exp(v_t) = \begin{pmatrix} \mathrm{Id}_{p-2} & 0 & 0 & 0 & 0 \\ 0 & 1 & v^t & q_J(v) & 0 \\ 0 & 0 & \mathrm{Id}_{q-p+2} & Jv & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{Id}_{p-2} \end{pmatrix}. \qquad \square$$

In particular, we deduce from [39, Proposition 7.4]:

Corollary 10.4 *Let $\rho\colon \Gamma_g \to \mathsf{SO}(p,q)$ be Θ–positive Anosov. For every $1 \le k \le p-2$ the image of $\xi^k_\rho(\partial\Gamma)$ is a C^1 submanifold of $\mathsf{Is}_k(\mathbb{R}^{p,q})$.*

Proof Since $\bigwedge^k \rho$ is $(1,1,2)$–hyperconvex, by [39, Proposition 7.4] its limit set is a C^1 submanifold of $\mathbb{P}(\bigwedge^k \mathbb{R}^{p,q})$. Since the inclusion $\bigwedge^k : \mathsf{Is}_k(\mathbb{R}^{p,q}) \to \mathbb{P}(\bigwedge^k \mathbb{R}^{p,q})$ is analytic, and the limit set of $\bigwedge^k \rho$ is the image under this inclusion of the limit set of ρ, the result follows. $\square$

We now turn to the proof of the last statement in Theorem D. Instead of directly verifying that the map ξ^{p-1}_ρ has Lipschitz image, we will study properties of the map $\xi^{\Theta_0}_\rho : \partial\Gamma_g \to \mathsf{G}/P_{\Theta_0}$, where

$$\Theta_0 = \{\mathsf{a}_{p-2}, \mathsf{a}_{p-1}\}.$$

The flag manifold G/P_{Θ_0} consists of nested pairs of isotropic subspaces of dimension $p-2$ and $p-1$.

Proposition 10.5 *Let $\rho\colon \Gamma_g \to \mathsf{SO}(p,q)$ be Θ–positive Anosov. The image of the map $\xi^{\Theta_0}_\rho : \partial\Gamma_g \to \mathsf{G}/P_{\Theta_0}$ is a Lipschitz submanifold of G/P_{Θ_0}.*

Proof Fix a point $z \in \partial\Gamma$ and assume without loss of generality that $\xi^k_\rho(z) = \langle e_1, \dots, e_k\rangle$. We denote by $\mathcal{A} \subset G/P_{\Theta_0}$ the set of points transverse to $\xi^{p-2,p-1}_\rho(z)$. We will show that the image of $\xi^{\Theta_0}_\rho|_{\partial\Gamma\setminus\{z\}}$ is a Lipschitz submanifold of $\mathcal{A}$. Denote by $\mathcal{A}_{p-2} \subset G/P_{\mathsf{a}_{p-2}}$ the set of isotropic subspaces of dimension $p-2$ transverse to $\xi^{p-2}_\rho(z) = \langle e_1, \dots, e_{p-2}\rangle$, by Z_{p-1} the $(p-1)$–isotropic subspace $Z_{p-1} := \xi^{p-1}_\rho(z) = \langle e_1, \dots, e_{p-1}\rangle$ and by $Z^\perp_{p-1}$ its orthogonal with respect to the form Q defining $\mathsf{SO}(p,q)$. Observe that we have a smooth map

$$\mathcal{A} \mapsto \mathcal{A}_{p-2} \times \mathsf{Is}_1(Z^\perp_{p-2}/Z_{p-2}), \quad (Y_{p-2}, Y_{p-1}) \mapsto (Y_{p-2}, [Y_{p-1} \cap Z^\perp_{p-2}]),$$

whose image is the product of $\mathcal{A}_{p-2}$ with the set $\mathcal{I}_Z$ of isotropic lines transverse to the image of $Z^\perp_{p-1}$. Indeed, for every pair $(Y_{p-2}, v) \in \mathcal{A}_{p-2} \times \mathcal{I}_Z$, the subspace $v + Z_{p-2}$ has dimension $p-1$ and $\dim((v + Z_{p-2}) \cap Y^\perp_{p-2}) = 1$ as $Y^\perp_{p-2}$ and Z_{p-2} are transverse. We then have $Y_{p-1} = Y_{p-2} + ((v + Z_{p-2}) \cap Y_{p-2})$.

Denote by $\xi_Z : \partial\Gamma\setminus\{z\} \to \mathcal{I}_Z$ the composition of the map $\xi^{p-2,p-1}$ and the projection to the second factor in the product decomposition. The form Q induces a form of signature $(2, q-p+2)$ on $Z^\perp_{p-2}/Z_{p-2}$, which gives rise to the notion of positive curves (as introduced in Section 9). We claim that ξ_Z is a positive curve. This amounts to showing that if $(x,y,z)\in\partial\Gamma$ is positively oriented then $\xi_Z(y)=s^Z\xi_Z(x)$ for some positive element s^Z in the unipotent radical of the stabilizer of $[Z_{p-1}]\in\mathsf{Is}_1(Z_{p-2}^\perp/Z_{p-2})$. Since the representation ρ is Θ–positive we know that $\xi(y)=s\cdot\xi(x)$ for some element in the positive semigroup U_Θ^+ and, as in the proof of Theorem 10.3, we can write $s=\exp(v_1)\cdots\exp(v_l)$ with $v_t\in c^0_{\mathsf{a}_{i_t}}$. Observe that, for every $v_t\in c^0_{\beta_{i_t}}$, $\exp(v_t)$ induces an element $\exp(v_t)^Z$ in the unipotent radical of the stabilizer of $[Z_{p-1}]\in\mathsf{Is}_1(Z_{p-2}^\perp/Z_{p-2})$, and the element $\exp(v_t)^Z$ is trivial unless $\beta_{i_t}=\mathsf{a}_{p-1}$, in which case $\exp(v_t)^Z$ belongs to the positive semigroup of the unipotent radical of the stabilizer of $[Z_{p-1}]$. As at least one of the v_t in the decomposition of s belongs to such a subgroup, we deduce that ξ_Z is positive, as we claimed. It follows from Proposition 9.3 that $\xi_Z(\partial\Gamma\setminus\{z\})$ is a Lipschitz submanifold of $\mathsf{Is}_1(Z_{p-2}^\perp/Z_{p-2})$.

As we know from Theorem 10.3 that ξ^{p-2} is a C^1–curve, we deduce that the curve $\xi^{p-2,p-1}$ is Lipschitz, being the image of a monotone map between a C^1–submanifold and a Lipschitz submanifold. □

10.1 The critical exponent on the symmetric space is rigid

Let $\iota_{2p-1}\colon\mathsf{PO}(1,2)\to\mathsf{PO}(p,p-1)\to\mathsf{PO}(p,q)$ be the composition of the irreducible representation of dimension $2p-1$ with the standard embedding of $\mathsf{PO}(p,p-1)$ into $\mathsf{PO}(p,q)$. We call any representation $\rho\colon\Gamma\to\mathsf{PO}(p,q)$ which is the composition of a Fuchsian representation with ι_{2p-1} a $(p,p-1)$–*Fuchsian* representation.

Lemma 10.6 *Let $\rho\colon\Gamma_g\to\mathsf{PO}(p,q)$ be Θ–positive Anosov. The barycenter of the affine simplex in E_Θ^* determined by $\{\mathsf{a}_1,\dots,\mathsf{a}_{p-2},\varepsilon_{p-1}\}$ belongs to $\mathcal{D}_{\rho(\Gamma),\Theta}$.*

Proof Recall that, in the case of Θ–positive representations in $\mathsf{PO}(p,q)$, the Levi–Anosov subspace is $\mathsf{E}_\Theta:=\ker(\mathsf{a}_p)$. In particular, for every $k\le p-2$ we have that a_k belongs to the dual of E_Θ, and belongs to the boundary of $\mathcal{D}_{\rho(\Gamma),\Theta}$ by Corollary 10.4. Furthermore $\varepsilon_{p-1}=\mathsf{a}_{p-1}+\mathsf{a}_p$ belongs to $\mathcal{D}_{\rho(\Gamma),\Theta}$, being the sum of a linear form with entropy one (the form a_{p-1} has entropy one by Proposition 10.5) and a linear form positive on the Weyl chamber (the root a_p). In particular, the form corresponding to the barycenter of the affine simplex they determine in E_Θ^* belongs to $\mathcal{D}_{\rho(\Gamma),\Theta}$. □

Theorem 10.7 *Let Γ be the fundamental group of a surface and let $\rho\colon\Gamma\to\mathsf{PO}(p,q)$ be Θ–positive Anosov. Then $h_\rho^{\mathcal{X}}\le h_{\rho_0}^{\mathcal{X}}$ for any $(p,p-1)$–Fuchsian representation ρ_0.*

If equality is achieved at a totally reducible representation η, then η splits as $W\oplus V$, where

(1) *W has signature $(p, p-1)$ and $\eta|_W$ has Zariski closure the irreducible* $\mathsf{PO}(2,1)$ *in* $\mathsf{PO}(p, p-1)$*,*

(2) *$\eta|_V$ lies in a compact group.*

Proof The inequality follows from Lemma 10.6, together with the convexity of $\mathcal{D}_{\rho(\Gamma),\Theta}$, established by Theorem 5.12.

Assume now that η is a totally reducible representation such that equality holds. We can assume that $p \geq 3$, as the result for $p = 2$ was proven by Collier, Tholozan and Toulisse [15, Theorem 4].

Let $\mathsf{G} = \overline{\eta(\Gamma)}^Z$ be the Zariski closure. By definition, G is a real reductive group. We consider G as an abstract group, and denote by $\Lambda\colon \mathsf{G} \to \mathsf{SO}(p,q)$ the inclusion representation and by

$$\phi\colon \mathfrak{g} \to \mathfrak{so}(p,q)$$

the associated Lie algebra morphism. Denote by $\mathfrak{a}_\mathsf{G}$ a Cartan subspace of $\mathfrak{g}$.

Since $h_\eta^{\mathcal{X}}$ attains it maximal value, Theorem 5.12 forces the Quint indicator set $\mathcal{Q}_{\eta(\Gamma),\Theta}$ to be the affine hyperplane of $(\mathsf{E}_\Theta)^*$ spanned by Δ. The strict convexity guaranteed by Theorem 5.12 implies that G has real rank at most 2. Moreover, we have that $\phi(\mathfrak{a}_\mathsf{G}) = \langle (2(p-1), 2(p-2), \dots, 2, 0), (0, \dots, 0, 1)\rangle$.

Denote by $T = \langle \xi_\eta^1(\partial\Gamma)\rangle$ the vector space spanned by the projective limit curve of η. Since η is totally reducible, the action of $\eta(\Gamma)$, and hence that of G, on T is irreducible.

Fix a Weyl chamber $\mathfrak{a}_\mathsf{G}^+$ and let $\chi \in \mathfrak{a}_\mathsf{G}^*$ be the highest weight of $\phi(\mathfrak{g})|_T$. Since η is a_1–Anosov, the attracting eigenvector of every element in $\eta(\Gamma)$, and hence of every purely loxodromic element of G, is in V. We therefore conclude that, for every $a \in \mathfrak{a}_\mathsf{G}^+$,

$$\chi(a) = \lambda_1(\phi(a)).$$

We denote by $\mathcal{L}_\eta^\mathsf{G} \subset \mathfrak{a}_\mathsf{G}^+$ Benoist's limit cone of $\eta(\Gamma)$ in G. As the representation η is a_2–Anosov, and thus $\mathcal{L}_\eta^\mathsf{G}$ avoids the only wall not orthogonal to the kernel of a_1, there exists a linear form $\mu \in \mathfrak{a}_\mathsf{G}^*$ such that, for every $a \in \mathcal{L}_\eta^\mathsf{G}$,

$$\mu(a) = \mathsf{a}_1(\lambda(\phi(a))).$$

Furthermore, as η is $(1,1,2)$–hyperconvex, for every $x \in \partial\Gamma$ the 2–dimensional space $\xi^{\mathsf{a}_2}(x)$ lies in T, and therefore $(\chi - \mu)(a) = \lambda_2(\phi(a))$, which implies that μ is a simple root and $\chi = (p-1)\mu$.

For a weight ψ of the representation $\phi(\mathfrak{g})|_T$, or of an irreducible factor of $\phi(\mathfrak{g})|_{T^\perp}$, denote by V^ψ the associated weight space. We obtain from the description of $\phi(\mathfrak{a}_\mathsf{G})$ that the weight spaces $V^{\chi-i\mu}$ for $i\in[\![0,2p-2]\!]$ are also 1–dimensional and contained in T. The weight space decomposition of T thus has the form

$$T=\bigoplus_{i=0}^{2p-2}V^{\chi-i\mu}\oplus V^0\oplus V^q\oplus V^{-q},$$

where V^0 consists of vectors in the kernel of $\phi(\mathfrak{a}_\mathsf{G}^+)$ (except $V^{\chi-(p-1)\mu}$) and V^q corresponds to the eigenvalue $\varepsilon_p(\lambda(\phi(a)))$. Here V^0 as well as V^q and V^{-q} could be instead contained in $T^\perp$, and therefore not appear in the decomposition.

Now let W denote the Weyl group of $\mathfrak{g}$. As the weight lattice of $\eta|_T$ is W–invariant, and there is no other weight of $\eta|_T$ at distance $p-1$ from the origin, we deduce that W is reducible, and $\mathfrak{g}$ splits as $\mathfrak{g}_1+\mathfrak{g}_2$. If μ is the root associated to $\mathfrak{g}_1$, we deduce from the fact that $V^{\chi-\mu}$, and thus $\mathfrak{g}_\mu$, is one-dimensional that $\mathfrak{g}_1=\mathfrak{sl}(2,\mathbb{R})$. As the actions of $\mathfrak{g}_1$ and $\mathfrak{g}_2$ commute and the highest weight space for the restricted action of $\mathfrak{g}_1$ is one-dimensional, we furthermore deduce that $\mathfrak{g}_2$ acts trivially on T. In particular, T is an irreducible $\mathfrak{sl}(2,\mathbb{R})$–module of dimension $2p-1$ and the signature of $T^\perp$ of the (p,q)–quadratic form preserved by $\mathfrak{so}(p,q)$ is thus either negative or $(1,q-p)$. In the first case we conclude that $\phi(\mathfrak{g})|_{T^\perp}$ is compact, which is the desired result.

In order to conclude the proof we need to exclude the second case. We know from Theorem 10.3 that for every $1\le k\le p-2$ and for every distinct $x,y,z\in\partial\Gamma$ the sum

$$\xi^k(x)+(\xi^k(y)\cap\xi^{d-k+1}(z))+\xi^{d-k-1}(z)$$

is direct. With an inductive argument we deduce that, for every $1\le k\le p-2$ and for every $\gamma\in\Gamma$, the k^{th} eigenline belongs to T and therefore the Anosov map ξ would be the boundary of a Fuchsian representation composed with an embedding of $\mathsf{PO}(1,2)\to\mathsf{PO}(p-1,p)\to\mathsf{PO}(p,q)$. However, such an embedding can never be positive because it has noncompact centralizer (compare with Remark 10.2). $\square$

References

[1] **T Barbot**, *Three-dimensional Anosov flag manifolds*, Geom. Topol. 14 (2010) 153–191 MR Zbl

[2] **T Barbot**, *Deformations of Fuchsian AdS representations are quasi-Fuchsian*, J. Differential Geom. 101 (2015) 1–46 MR Zbl

[3] **T Barbot**, **Q Mérigot**, *Anosov AdS representations are quasi-Fuchsian*, Groups Geom. Dyn. 6 (2012) 441–483 MR Zbl

[4] **Y Benoist**, *Propriétés asymptotiques des groupes linéaires*, Geom. Funct. Anal. 7 (1997) 1–47 MR Zbl

[5] **Y Benoist**, *Convexes divisibles, I*, from "Algebraic groups and arithmetic" (S G Dani, G Prasad, editors), Tata Inst. Fund. Res., Mumbai (2004) 339–374 MR Zbl

[6] **J Bochi**, **R Potrie**, **A Sambarino**, *Anosov representations and dominated splittings*, J. Eur. Math. Soc. 21 (2019) 3343–3414 MR Zbl

[7] **R Bowen**, *Hausdorff dimension of quasicircles*, Inst. Hautes Études Sci. Publ. Math. 50 (1979) 11–25 MR Zbl

[8] **S B Bradlow**, **O García-Prada**, **P B Gothen**, *Deformations of maximal representations in* $\mathrm{Sp}(4,\mathbb{R})$, Q. J. Math. 63 (2012) 795–843 MR Zbl

[9] **M Bridgeman**, **R Canary**, **F Labourie**, **A Sambarino**, *The pressure metric for Anosov representations*, Geom. Funct. Anal. 25 (2015) 1089–1179 MR Zbl

[10] **M Burger**, **A Iozzi**, **F Labourie**, **A Wienhard**, *Maximal representations of surface groups: symplectic Anosov structures*, Pure Appl. Math. Q. 1 (2005) 543–590 MR Zbl

[11] **M Burger**, **A Iozzi**, **A Wienhard**, *Tight homomorphisms and Hermitian symmetric spaces*, Geom. Funct. Anal. 19 (2009) 678–721 MR Zbl

[12] **M Burger**, **A Iozzi**, **A Wienhard**, *Surface group representations with maximal Toledo invariant*, Ann. of Math. 172 (2010) 517–566 MR Zbl

[13] **L Carvajales**, *Counting problems for special-orthogonal Anosov representations*, Ann. Inst. Fourier (Grenoble) 70 (2020) 1199–1257 MR Zbl

[14] **J L Clerc**, **B Ørsted**, *The Maslov index revisited*, Transform. Groups 6 (2001) 303–320 MR Zbl

[15] **B Collier**, **N Tholozan**, **J Toulisse**, *The geometry of maximal representations of surface groups into* $\mathrm{SO}_0(2,n)$, Duke Math. J. 168 (2019) 2873–2949 MR Zbl

[16] **J Danciger**, **F Guéritaud**, **F Kassel**, *Convex cocompactness in pseudo-Riemannian hyperbolic spaces*, Geom. Dedicata 192 (2018) 87–126 MR Zbl

[17] **J Danciger**, **F Guéritaud**, **F Kassel**, *Proper affine actions for right-angled Coxeter groups*, Duke Math. J. 169 (2020) 2231–2280 MR Zbl

[18] **S Dey**, **M Kapovich**, *Patterson–Sullivan theory for Anosov subgroups*, Trans. Amer. Math. Soc. 375 (2022) 8687–8737 MR Zbl

[19] **A Douady**, **J Oesterlé**, *Dimension de Hausdorff des attracteurs*, C. R. Acad. Sci. Paris Sér. A-B 290 (1980) 1135–1138 MR Zbl

[20] **L C Evans**, **R F Gariepy**, *Measure theory and fine properties of functions*, CRC, Boca Raton, FL (2015) MR Zbl

[21] **K J Falconer**, *The Hausdorff dimension of self-affine fractals*, Math. Proc. Cambridge Philos. Soc. 103 (1988) 339–350 MR Zbl

[22] **H Federer**, *Geometric measure theory*, Grundl. Math. Wissen. 153, Springer (1969) MR Zbl

[23] **O Glorieux**, **D Monclair**, *Regularity of limit sets of AdS quasi-Fuchsian groups*, preprint (2018) arXiv 1809.10639

[24] **O Glorieux**, **D Monclair**, *Critical exponent and Hausdorff dimension in pseudo-Riemannian hyperbolic geometry*, Int. Math. Res. Not. 2021 (2021) 13661–13729 MR Zbl

[25] **O Glorieux**, **D Monclair**, **N Tholozan**, *Hausdorff dimension of limit sets for projective Anosov representations*, preprint (2019) arXiv 1902.01844

[26] **O Guichard**, **A Wienhard**, *Topological invariants of Anosov representations*, J. Topol. 3 (2010) 578–642 MR Zbl

[27] **O Guichard**, **A Wienhard**, *Positivity and higher Teichmüller theory*, from "European Congress of Mathematics" (V Mehrmann, M Skutella, editors), Eur. Math. Soc., Zürich (2018) 289–310 MR Zbl

[28] **O Hamlet**, **M B Pozzetti**, *Classification of tight homomorphisms*, preprint (2014) arXiv 1412.6398

[29] **Harish-Chandra**, *Representations of semisimple Lie groups, VI: Integrable and square-integrable representations*, Amer. J. Math. 78 (1956) 564–628 MR Zbl

[30] **J E Humphreys**, *Linear algebraic groups*, Graduate Texts in Math. 21, Springer (1975) MR Zbl

[31] **J L Kaplan**, **J A Yorke**, *Chaotic behavior of multidimensional difference equations*, from "Functional differential equations and approximation of fixed points" (H-O Peitgen, H-O Walther, editors), Lecture Notes in Math. 730, Springer (1979) 204–227 MR Zbl

[32] **I Kapovich**, **N Benakli**, *Boundaries of hyperbolic groups*, from "Combinatorial and geometric group theory" (S Cleary, R Gilman, A G Myasnikov, V Shpilrain, editors), Contemp. Math. 296, Amer. Math. Soc., Providence, RI (2002) 39–93 MR Zbl

[33] **M Kapovich**, **B Leeb**, **J Porti**, *Morse actions of discrete groups on symmetric space*, preprint (2014) arXiv 1403.7671

[34] **M Kapovich**, **B Leeb**, **J Porti**, *A Morse lemma for quasigeodesics in symmetric spaces and Euclidean buildings*, Geom. Topol. 22 (2018) 3827–3923 MR Zbl

[35] **F Labourie**, *Anosov flows, surface groups and curves in projective space*, Invent. Math. 165 (2006) 51–114 MR Zbl

[36] **F Ledrappier**, *Structure au bord des variétés à courbure négative*, Sémin. Théor. Spectr. Géom. 13 (1995) 97–122 MR Zbl

[37] **G Mess**, *Lorentz spacetimes of constant curvature*, Geom. Dedicata 126 (2007) 3–45 MR Zbl

[38] **R Potrie**, **A Sambarino**, *Eigenvalues and entropy of a Hitchin representation*, Invent. Math. 209 (2017) 885–925 MR Zbl

[39] **M B Pozzetti**, **A Sambarino**, **A Wienhard**, *Conformality for a robust class of non-conformal attractors*, J. Reine Angew. Math. 774 (2021) 1–51 MR Zbl

[40] **J-F Quint**, *Cônes limites des sous-groupes discrets des groupes réductifs sur un corps local*, Transform. Groups 7 (2002) 247–266 MR Zbl

[41] **J-F Quint**, *Divergence exponentielle des sous-groupes discrets en rang supérieur*, Comment. Math. Helv. 77 (2002) 563–608 MR Zbl

[42] **J-F Quint**, *Mesures de Patterson–Sullivan en rang supérieur*, Geom. Funct. Anal. 12 (2002) 776–809 MR Zbl

[43] **J-F Quint**, *An overview of Patterson–Sullivan theory* (2006) Available at `http://www.math.u-bordeaux.fr/~jquint/publications/courszurich.pdf`

[44] **A Sambarino**, *Hyperconvex representations and exponential growth*, Ergodic Theory Dynam. Systems 34 (2014) 986–1010 MR Zbl

[45] **A Sambarino**, *The orbital counting problem for hyperconvex representations*, Ann. Inst. Fourier (Grenoble) 65 (2015) 1755–1797 MR Zbl

[46] **D Sullivan**, *The density at infinity of a discrete group of hyperbolic motions*, Inst. Hautes Études Sci. Publ. Math. 50 (1979) 171–202 MR Zbl

[47] **J Tits**, *Représentations linéaires irréductibles d'un groupe réductif sur un corps quelconque*, J. Reine Angew. Math. 247 (1971) 196–220 MR Zbl

[48] **T Zhang**, **A Zimmer**, *Regularity of limit sets of Anosov representations*, preprint (2019) arXiv 1903.11021

Mathematical Institute, Heidelberg University
Heidelberg, Germany

IMJ-PRG, Sorbonne Université, CNRS
Paris, France

Mathematisches Institut, Universität Heidelberg
Heidelberg, Germany

`pozzetti@mathi.uni-heidelberg.de`, `andres.sambarino@imj-prg.fr`,
`wienhard@uni-heidelberg.de`

Proposed: David Fisher Received: 25 June 2021
Seconded: Mladen Bestvina, Ian Agol Revised: 31 December 2021

Geometry & Topology **27**:8 (2023) 3361–3385
DOI: 10.2140/gt.2023.27.3361
Published: 9 November 2023

The deformation space of geodesic triangulations and generalized Tutte's embedding theorem

YANWEN LUO
TIANQI WU
XIAOPING ZHU

We prove the contractibility of the deformation space of the geodesic triangulations on a closed surface of negative curvature. This solves an open problem, proposed by Connelly, Henderson, Ho and Starbird (1983), in the case of hyperbolic surfaces. The main part of the proof is a generalization of Tutte's embedding theorem for closed surfaces of negative curvature.

54C25, 55Q52, 57N65, 57S05, 58D10

1 Introduction

We study the deformation space of geodesic triangulations of a surface within a fixed homotopy class. Such a space can be viewed as a discrete analogue of the space of surface diffeomorphisms homotopic to the identity. Our main theorem is:

Theorem 1.1 *For a closed orientable surface of negative curvature, the space of geodesic triangulations in a homotopy class is contractible. In particular, it is connected.*

The group of diffeomorphisms of a smooth surface is a fundamental object in the study of low-dimensional topology. Determining the homotopy types of diffeomorphism groups has profound implications for a wide range of problems in Teichmüller spaces, mapping class groups, and geometry and topology of 3–manifolds. Smale [23] proved that the group of diffeomorphisms of a closed 2–disk which pointwise fix the boundary is contractible. This enabled him to show that the group of orientation-preserving diffeomorphisms of the 2–sphere is homotopy equivalent to SO(3) [23]. Earle and Eells [10] identified the homotopy type of the group of diffeomorphisms homotopic to the identity for any closed surface. In particular, this topological group is contractible for a closed orientable surface with genus greater than one. It is consistent with our Theorem 1.1 for the discrete analogue.

Cairns [6] initiated the investigation of the topology of the space of geodesic triangulations and proved that, if the surface is a geometric triangle in the Euclidean plane, the space of geodesic triangulations with fixed boundary edges is connected. A series of further developments culminated in a discrete version of Smale's theorem, proved by Bloch, Connelly and Henderson [2]:

Theorem 1.2 *The space of geodesic triangulations of a convex polygon with fixed boundary edges is homeomorphic to a Euclidean space. In particular, it is contractible.*

A simple proof of the contractibility of the space above is provided in Luo [21] using Tutte's embedding theorem [24]. It also provides examples showing that the homotopy type of this space can be complicated if the boundary of the polygon is not convex. For closed surfaces it is conjectured in Connelly, Henderson, Ho and Starbird [9] that:

Conjecture 1.3 *The space of geodesic triangulations of a closed orientable surface with constant curvature deformation retracts to the group of isometries of the surface homotopic to the identity.*

The connectivity of these spaces has been explored by Cairns [6], Chambers, Erickson, Lin and Parsa [7] and Hass and Scott [18]. Awartani and Henderson [1] identified a contractible subspace in the space of geodesic triangulations of the 2–sphere. Hass and Scott [18] showed that the space of geodesic triangulation of a surface with a hyperbolic metric is contractible if the triangulation contains only one vertex. Recently, the authors [22] and Erickson and Lin [11] proved this conjecture independently in the case of flat tori. Our main result affirms Conjecture 1.3 in the case of hyperbolic surfaces and generalizes its conclusion to surfaces of negative curvatures.

One practical application of our work concerns the graph morphing on higher-genus surfaces. Computing morphs between graphs has a wide range of applications in geometric comparison, animation, and modeling. The 1–skeleton of geodesic triangular mesh is one of the most common graphs on a surface. As a fundamental result, our main theorem implies that, on a closed surface of negative curvature, any two geodesic triangular meshes can be morphed to each other if they have the same combinatorial structure. Furthermore, in the proof of our main theorem, we generalize Tutte's embedding theorem to higher-genus surfaces. Following the idea initiated by Floater and Gotsman [14], we can explicitly construct such morphs by linearly interpolating the nonsymmetric edge weights. A similar idea has been applied for graph morphing on flat tori in work by Chambers, Erickson, Lin and Parsa [7] and Erickson and Lin [11].

1.1 Setup and the main theorem

Assume M is a connected closed orientable smooth surface with a smooth Riemannian metric g of nonpositive Gaussian curvature. A topological triangulation of M can be identified as a homeomorphism ψ from $|T|$ to M, where $|T|$ is the carrier of a 2–dimensional simplicial complex $T = (V, E, F)$ with the vertex set V, the edge set E, and the face set F. For convenience, we label the vertices as $1, 2, \dots, n$, where $n = |V|$ is the number of vertices. The edge in E determined by vertices i and j is written ij. Each edge is identified with the closed unit interval $[0, 1]$.

Let $T^{(1)}$ be the 1–skeleton of T, and denote by $X = X(M, T, \psi)$ the space of geodesic triangulations homotopic to $\psi|_{T^{(1)}}$. More specifically, X contains all the embeddings $\varphi\colon T^{(1)} \to M$ such that

(i) the restriction φ_{ij} of φ to the edge ij is a geodesic parametrized with constant speed, and

(ii) φ is homotopic to $\psi|_{T^{(1)}}$.

Given an embedding φ in X, φ_{ij} is often identified as a map from $[0, 1]$ to M such that $\varphi(0) = i$, $\varphi(1) = j$ and $\varphi_{ij}(t)$ represents the point on the edge ij that is t along the geodesic from i to j parametrized on $[0, 1]$.

It has been proved by Colin de Verdière [8] that such $X(M, T, \psi)$ is always nonempty. Further, X is naturally a metric space, with the distance function

$$d_X(\varphi, \phi) = \max_x d_g(\varphi(x), \phi(x)).$$

Then our main theorem is formally stated as follows:

Theorem 1.4 *If (M, g) has strictly negative Gaussian curvature, then $X(M, T, \psi)$ is contractible. In particular, it is connected.*

Here we consider only surfaces of negative curvature since this ensures the uniqueness of the geodesic in a homotopy class, and our estimates using the $\mathrm{CAT}(k)$ comparison theorems of triangles rely on a strictly negative upper bound of the curvature of the surface.

1.2 Generalized Tutte's embedding

Let $\widetilde{X} = \widetilde{X}(M, T, \psi)$ be the superspace of X containing all the continuous maps $\varphi\colon T^{(1)} \to M$ satisfying that

(i) the restriction φ_{ij} of φ to the edge ij is a geodesic parametrized with constant speed, and

(ii) φ is homotopic to $\psi|_{T^{(1)}}$.

The key difference between X and $\widetilde{X}$ is that elements in $\widetilde{X}$ may not be embeddings of $T^{(1)}$ to M. The space $\widetilde{X}$ is also naturally a metric space, with the same distance function

$$d_{\widetilde{X}}(\varphi, \phi) = \max_x d_g(\varphi(x), \phi(x)).$$

We call an element in $\widetilde{X}$ a *geodesic mapping*. A geodesic mapping is determined by the positions $q_i = \varphi(i)$ of the vertices and the homotopy classes of φ_{ij} relative to the endpoints q_i and q_j. In particular, this holds for geodesic triangulations. Since we can perturb the vertices of a geodesic triangulation to generate another, X is a $2n$–dimensional manifold.

Let (i, j) be the directed edge starting from the vertex i and ending at the vertex j. Denote by $\vec{E} = \{(i, j) : ij \in E\}$ the set of directed edges of T. A positive vector $w \in \mathbb{R}^{\vec{E}}_{>0}$ is called a *weight* of T. For any weight w and geodesic mapping $\varphi \in \widetilde{X}$, we say φ is *w–balanced* if, for any $i \in V$,

$$\sum_{j:ij\in E} w_{ij} v_{ij} = 0.$$

Here $v_{ij} \in T_{q_i}M$ is defined with the exponential map $\exp\colon TM \to M$ such that $\exp_{q_i}(t v_{ij}) = \varphi_{ij}(t)$ for $t \in [0, 1]$.

The main part of the proof of Theorem 1.4 is to generalize Tutte's embedding theorem (see Theorem 9.2 in [24] or Theorem 6.1 in Floater [13]) to closed surfaces of negative curvature. Specifically, we prove the following two theorems:

Theorem 1.5 *Assume (M, g) has strictly negative Gaussian curvature. For any weight w there exists a unique geodesic mapping $\varphi \in \widetilde{X}(M, T, \psi)$ that is w–balanced. The induced map $\Phi(w) = \varphi$ is continuous from $\mathbb{R}^{\vec{E}}_{>0}$ to $\widetilde{X}$.*

Theorem 1.6 *If $\varphi \in \widetilde{X}$ is w–balanced for some weight w, then $\varphi \in X$.*

Theorem 1.6 can be regarded as a generalization of the embedding theorems of Colin de Verdière (see Theorem 2 in [8]) and Hass and Scott (see Lemma 10.12 in [18]), which imply that the minimizer of the discrete Dirichlet energy

$$E(\varphi) = \frac{1}{2}\sum_{ij\in E} w_{ij} l_{ij}^2$$

among the maps φ in the homotopy class of $\psi|_{T^{(1)}}$ is a geodesic triangulation. Here l_{ij} is the geodesic length of φ_{ij} in M. The minimizer is a w–balanced geodesic mapping with $w_{ij} = w_{ji}$ for $ij \in E$. Hence, Theorem 1.6 extends the previous results from the cases of symmetric weights to nonsymmetric weights. We believe that the proofs of Colin de Verdière [8] and Hass and Scott [18] could be easily modified to work with our nonsymmetric case. Nevertheless, we give a new proof in Section 3 to make the paper self-contained.

1.3 Mean value coordinates and the proof of Theorem 1.4

Theorems 1.5 and 1.6 give a continuous map Φ from $\mathbb{R}^{\vec{E}}_{>0}$ to X. To map a geodesic embedding to a weight, we use the *mean value coordinates* introduced by Floater [12]. Given $\varphi \in X$ the mean value coordinates are defined to be

$$w_{ij} = \frac{\tan\bigl(\frac{1}{2}\alpha_{ij}\bigr) + \tan\bigl(\frac{1}{2}\beta_{ij}\bigr)}{|v_{ij}|},$$

where $|v_{ij}|$ equals the geodesic length of $\varphi_{ij}([0,1])$, and α_{ij} and β_{ij} are the inner angles in $\varphi(T^{(1)})$ at $\varphi(i)$ sharing the edge $\varphi_{ij}([0,1])$. The construction of mean value coordinates gives a continuous map Ψ from X to $\mathbb{R}^{\vec{E}}_{>0}$. Further, by Floater's mean value theorem (see Proposition 1 in [12]), any $\varphi \in X$ is $\Psi(\varphi)$–balanced. Namely, $\Phi \circ \Psi = \mathrm{id}_X$. Then Theorem 1.4 is a direct consequence of Theorems 1.5 and 1.6.

Proof of Theorem 1.4 Since $\mathbb{R}^{\vec{E}}_{>0}$ is contractible, $\Psi \circ \Phi$ is homotopic to the identity map. Since $\Phi \circ \Psi = \mathrm{id}_X$, X is homotopy equivalent to the contractible space $\mathbb{R}^{\vec{E}}_{>0}$. □

We will prove Theorem 1.5 in Section 2 and Theorem 1.6 in Section 3.

Acknowledgment The authors were supported in part by NSF 1737876, NSF 1760471, NSF DMS FRG 1760527 and NSF DMS 1811878.

2 Proof of Theorem 1.5

Theorem 1.5 consists of three parts: the existence of the w–balanced geodesic mapping, the uniqueness of the w–balanced geodesic mapping and the continuity of the map Φ.

In this section we will first parametrize $\widetilde{X}$ by $\mathcal{M}$, where $\mathcal{M}$ is the product manifold of the n copies of the universal cover $\widetilde{M}$ of M. Then we prove the three parts in Sections 2.1, 2.2 and 2.3, respectively.

For the proof we mainly work on the universal covering space $\widetilde{M}$ instead of the original surface M. This is because a geodesic arc is uniquely determined by its endpoints in $\widetilde{M}$ but not in M, and thus the geodesic triangulation of M in the same homotopy class is naturally parametrized by the lifted vertices in $\widetilde{M}$. The condition of strictly negative curvature is mostly needed in the proof of the existence of the w–balanced mappings, where we frequently compare geodesic triangles in $\widetilde{M}$ with geodesic triangles of constant negative curvature.

Assume that p is the covering map from $\widetilde{M}$ to M, and Γ is the corresponding group of deck transformations of the covering so that $\widetilde{M}/\Gamma = M$. For any $i \in V$, fix a lifting $\tilde{q}_i \in \widetilde{M}$ of $q_i \in M$. For any edge ij, denote by $\tilde{\varphi}_{ij}(t)$ by the lifting of $\varphi_{ij}(t)$ such that $\tilde{\varphi}_{ij}(0) = \tilde{q}_i$. Here $\tilde{\varphi}_{ij}(1)$ may not be equal to $\tilde{q}_j$, but $p(\tilde{\varphi}_{ij}(1)) = \varphi_{ij}(1) = q_j = p(\tilde{q}_j)$, and so there exists a unique deck transformation $A_{ij} \in \Gamma$ such that $\tilde{\varphi}_{ij}(1) = A_{ij}\tilde{q}_j$. Notice that the deck transformation A_{ij} depends on the choice of the lifts $\tilde{q}_i$ and $\tilde{q}_j$ of q_i and q_j, respectively. We can deduce that $A_{ij} = A_{ji}^{-1}$ for any edge ij.

Equip $\widetilde{M}$ with the natural pullback Riemannian metric $\tilde{g}$ of g with negative Gaussian curvature. This metric is equivariant with respect to Γ. For any $x, y \in \widetilde{M}$, there exists a unique geodesic with constant speed parametrization $\gamma_{x,y}\colon [0,1] \to \widetilde{M}$ such that $\gamma_{x,y}(0) = x$ and $\gamma_{x,y}(1) = y$. We can naturally parametrize $\widetilde{X}$ as follows:

Theorem 2.1 *For any $(x_1, \dots, x_n) \in \mathcal{M}$, define $\varphi = \varphi[x_1, \dots, x_n]$ as*

$$\varphi_{ij}(t) = p \circ \gamma_{x_i, A_{ij}x_j}(t)$$

for any $ij \in E$ and $t \in [0,1]$. Then φ is a well-defined geodesic mapping in $\widetilde{X}$, and the map $(x_1, \dots, x_n) \mapsto \varphi[x_1, \dots, x_n]$ is a homeomorphism from $\mathcal{M}$ to $\widetilde{X}$.

We omit the proof of Theorem 2.1, which is routine but lengthy. In the remainder of this section, for any $x, y, z \in \widetilde{M}$ and $u, v \in T_x\widetilde{M}$:

(i) $d(x,y)$ is the intrinsic distance between x and y in $(\widetilde{M}, \tilde{g})$.

(ii) $v(x,y) = \exp_x^{-1} y \in T_x\widetilde{M}$.

(iii) $\triangle xyz$ is the geodesic triangle in $\widetilde{M}$ with vertices x, y and z, which could possibly be degenerate.

(iv) $\angle yxz$ is the inner angle of $\triangle xyz$ at x if $d(x,y) > 0$ and $d(x,z) > 0$.

(v) $|v|$ is the norm of v under the metric $\tilde{g}_x$.

(vi) $u \cdot v$ is the inner product of u and v under the metric $\tilde{g}_x$.

By scaling the metric if necessary, we may assume that the Gaussian curvatures of (M, g) and $(\widetilde{M}, \tilde{g})$ are bounded above by -1.

2.1 Uniqueness

We first prove Lemma 2.2 using CAT(0) geometry. See Theorem 4.3.5 in [4] and Theorem 1A.6 in [3] for the well-known comparison theorems.

Lemma 2.2 *Assume $x, y, z \in \widetilde{M}$. Then*

(i) $|v(z,x) - v(z,y)| \le d(x,y)$, *and*

(ii) $v(x,y) \cdot v(x,z) + v(y,x) \cdot v(y,z) \ge d(x,y)^2$,

and equality holds if and only if $\triangle xyz$ is degenerate.

Proof If $\triangle xyz$ is degenerate then there exists a geodesic γ in $\widetilde{M}$ such that $x, y, z \in \gamma$, and then the proof is straightforward, so we assume that $\triangle xyz$ is nondegenerate.

(i) Three points $v(z,x)$, $v(z,y)$, and 0 in $T_z\widetilde{M}$ determine a Euclidean triangle, where $|v(z,x)| = d(x,z)$, $|v(z,y)| = d(z,y)$ and the angle between $v(z,x)$ and $v(z,y)$ is equal to $\angle xzy$. Then, by the CAT(0) comparison theorem,

$$|v(z,x) - v(z,y)| < d(x,y).$$

(ii) Let $x', y', z' \in \mathbb{R}^2$ be such that

$$|x' - z'|_2 = |v(x,z)|, \quad |y' - z'|_2 = |v(y,z)| \quad \text{and} \quad |x' - y'|_2 = |v(x,y)|.$$

Then, by the CAT(0) comparison theorem, $\angle yxz < \angle y'x'z'$ and $\angle xyz < \angle x'y'z'$. Hence,

$$\begin{aligned} v(x,y) \cdot v(x,z) + v(y,x) \cdot v(y,z) &> (y' - x') \cdot (z' - x') + (x' - y') \cdot (z' - y') \\ &= |x' - y'|_2^2 = d(x,y)^2. \end{aligned}$$

□

Proof of uniqueness in Theorem 1.5 If φ is not unique, assume $\varphi[x_1, \dots, x_n]$ and $\varphi[x'_1, \dots, x'_n]$ are two different geodesic mappings that are both w–balanced for some weight w. We are going to prove a discrete maximum principle for the function $j \mapsto d(x_j, x'_j)$. Assume $i \in V$ is such that $d(x_i, x'_i) = \max_{j \in V} d(x_j, x'_j) > 0$. By lifting the w–balanced assumption to $\widetilde{M}$, we have that

$$\sum_{\{j : ij \in E\}} w_{ij} v(x_i, A_{ij} x_j) = 0, \tag{1}$$

and

$$\sum_{\{j:ij\in E\}} w_{ij}\, v(x_i', A_{ij}x_j') = 0. \tag{2}$$

Then, by Lemma 2.2(i) and (1),

$$\begin{aligned}\Big|\sum_{\{j:ij\in E\}} w_{ij}\, v(x_i, A_{ij}x_j')\Big| &= \Big|\sum_{\{j:ij\in E\}} w_{ij}\, v(x_i, A_{ij}x_j') - \sum_{\{j:ij\in E\}} w_{ij}\, v(x_i, A_{ij}x_j)\Big| \\ &\le \sum_{\{j:ij\in E\}} w_{ij}\, d(A_{ij}x_j, A_{ij}x_j') = \sum_{\{j:ij\in E\}} w_{ij}\, d(x_j, x_j') \\ &\le d(x_i, x_i') \sum_{\{j:ij\in E\}} w_{ij}.\end{aligned}$$

By part (ii) of Lemma 2.2, (2), and the Cauchy–Schwartz inequality,

$$\begin{aligned}&d(x_i, x_i')\cdot\Big|\sum_{\{j:ij\in E\}} w_{ij}\, v(x_i, A_{ij}x_j')\Big| \\ &\qquad \ge v(x_i, x_i')\cdot \sum_{\{j:ij\in E\}} w_{ij}\, v(x_i, A_{ij}x_j') + v(x_i', x_i)\cdot \sum_{\{j:ij\in E\}} w_{ij}\, v(x_i', A_{ij}x_j') \\ &\qquad \ge \sum_{\{j:ij\in E\}} w_{ij}\cdot d(x_i, x_i')^2.\end{aligned}$$

Therefore, equality holds in both inequalities above. Then, for any neighbor j of i, $d(x_j, x_j') = d(x_i, x_i') = \max_{k\in V} d(x_k, x_k')$, and $A_{ij}x_j$ is on the geodesic determined by x_i and x_i'. Hence, the one-ring neighborhood of $p(x_i)$ in $\varphi[x_1, \dots, x_n](T^{(1)})$ degenerates to a geodesic arc. By the connectedness of the surface we can repeat the above argument and deduce that $d(x_j, x_j') = d(x_i, x_i')$ for any $j \in V$. Further, $\varphi[x_1, \dots, x_n](\partial\sigma)$ degenerates to a geodesic arc for any triangle $\sigma \in F$.

It is not difficult to extend $\varphi[x_1, \dots, x_n]$ to a continuous map $\tilde\varphi$ from $|T|$ to M such that $\tilde\varphi(\partial\sigma) = \varphi[x_1, \dots, x_n](\partial\sigma)$ is the union of three geodesic arcs for any triangle $\sigma \in F$

It is also not difficult to prove that $\tilde\varphi$ is homotopic to ψ. Therefore, $\tilde\varphi$ is degree one and surjective. This contradicts that $\tilde\varphi(|T|)$ is a finite union of geodesic arcs. □

2.2 Existence

Here we prove a stronger existence result:

Theorem 2.3 *Given a compact subset K of $\mathbb{R}^{\vec{E}}_{>0}$ there exists a compact subset $K' = K'(M, T, \psi, K)$ of $\widetilde{X}$ such that, for any $w \in K$, there exists a w–balanced geodesic mapping $\varphi \in K'$.*

We first introduce the topological Lemma 2.4 and the key Lemma 2.5.

Lemma 2.4 *Suppose $B^n = \{x \in \mathbb{R}^n : |x| \le 1\}$ is the unit ball in $\mathbb{R}^n$, and $f : B^n \to \mathbb{R}^n$ is a continuous map such that $x \neq f(x)/|f(x)|$ for any $x \in \partial B^n = S^{n-1}$ with $f(x) \neq 0$. Then f has a zero in B^n.*

Proof If not, $g(x) = f(x)/|f(x)|$ is a continuous map from B^n to ∂B^n. Since B^n is contractible, $g(x)$ is nullhomotopic, and thus $g|_{S^{n-1}}$ is also nullhomotopic. Since $g(x) \neq x$, it is easy to verify that

$$H(x,t) = \frac{tg(x) + (1-t)(-x)}{|tg(x) + (1-t)(-x)|}$$

is a homotopy between $g|_{S^{n-1}}$ and $-\mathrm{id}\,|_{S^{n-1}}$. This contradicts that $-\mathrm{id}\,|_{S_{n-1}}$ is not nullhomotopic. □

Lemma 2.5 *Fix an arbitrary point $q \in \widetilde{M}$. If $w \in \mathbb{R}^{\vec{E}}_{>0}$ and $(x_1, \dots, x_n) \in \mathcal{M}$ satisfies*

$$v(x_i, q) \cdot \sum_{\{j : ij \in E\}} w_{ij} v(x_i, A_{ij} x_j) \le 0 \tag{3}$$

for any $i \in V$, then

$$\sum_{i \in V} d(x_i, q)^2 < R^2$$

for some constant $R > 0$ which depends only on M, T, ψ, q and

$$\lambda_w := \frac{\max_{ij \in E} w_{ij}}{\min_{ij \in E} w_{ij}}.$$

The vector in Figure 1,

$$r_i = \sum_{\{j : ij \in E\}} w_{ij} v(x_i, A_{ij} x_j),$$

is defined as the *residue vector* r_i at x_i of $\varphi[x_1, \dots, x_n]$ with respect to the weight w. Notice that a geodesic mapping φ is w–balanced if and only if all its residue vectors vanish with respect to w. Lemma 2.5 means that, if all the residue vectors are dragging the x_i away from q, then all the x_i must stay a bounded distance from q.

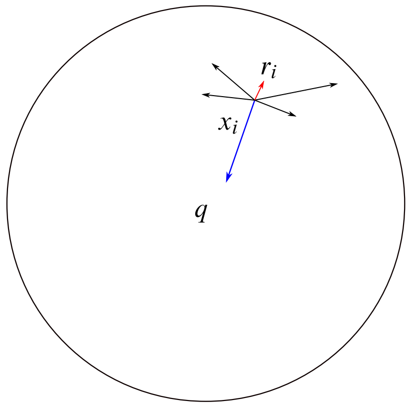

Figure 1: The residue vector and Lemma 2.5.

Our notion of w–balancedness is closely related to the *Riemannian center of mass* developed by Grove and Karcher [17]. In a w–balanced geodesic mapping, each point can be viewed as the weighted center of mass of its neighboring points. The defining formula of our residue vectors also appears in [5; 19]. A survey of Riemannian center of mass by Karcher can be found in [20]. The definition of a residue vector is also similar to the concept of a *discrete tension field* in [15].

Proof of Theorem 2.3 Fix an arbitrary basepoint $q \in \widetilde{M}$. Then by Lemma 2.5 we can pick a sufficiently large constant $R = R(M, T, \psi, K) > 0$ such that, if

$$\sum_{i=1}^{n} d(x_i, q)^2 = R^2,$$

there exists $i \in V$ such that

$$v(x_i, q) \cdot \sum_{\{j : ij \in E\}} w_{ij} v(x_i, A_{ij} x_j) > 0.$$

We will prove that the compact set

$$K' = \left\{ \varphi[x_1, \dots, x_n] \;\middle|\; \sum_{i=1}^{n} d(x_i, q)^2 \le R^2 \right\}$$

is satisfactory.

For $x \in \widetilde{M}$ let $P_x : T_x \widetilde{M} \to T_q \widetilde{M}$ be the parallel transport along the geodesic $\gamma_{x,q}$. Set

$$B = \left\{ (v_1, \dots, v_n) \in (T_q \widetilde{M})^n \;\middle|\; \sum_{i=1}^{n} |v_i|^2 \le 1 \right\},$$

a Euclidean $2n$–dimensional unit ball, and construct a map $F\colon B \to (T_q\widetilde{M})^n$ in three steps: Firstly, we define n points $x_1, \dots, x_n \in \widetilde{M}$ by $x_i(v_1, \dots, v_n) = \exp_q(Rv_i)$. Secondly, we compute the residue vector at each x_i as

$$r_i = \sum_{\{j:ij\in E\}} w_{ij} v(x_i, A_{ij}x_j) \in T_{x_i}\widetilde{M}.$$

Lastly, we pull back the residues to $T_q\widetilde{M}$ via $F(v_1, \dots, v_n) = (P_{x_1}(r_1), \dots, P_{x_n}(r_n))$.

Notice that the map $(v_1, \dots, v_n) \mapsto \varphi[x_1, \dots, x_n]$ is a homeomorphism from B to K', and $F(v_1, \dots, v_n) = 0$ if and only if the corresponding $\varphi[x_1, \dots, x_n]$ in K' is a w–balanced map. Hence, it suffices to prove that F has a zero in B. By Lemma 2.4, it suffices to prove that, for any $(v_1, \dots, v_n) \in \partial B$,

$$(v_1, \dots, v_n) \neq \frac{F(v_1, .., v_n)}{|F(v_1, \dots, v_n)|}.$$

Suppose $(v_1, \dots, v_n)$ is an arbitrary point on ∂B. Then it suffices to prove that there exists $i \in V$ such that $v_i \cdot F_i(v_1, \dots, v_n) = v_i \cdot P_{x_i}(r_i) < 0$.

Notice that $x_1(v_1, \dots, v_n), \dots, x_n(v_1, \dots, v_n)$ satisfy that $\sum_{i=1}^n d(q, x_i)^2 = R^2$, so, by our assumption on R, there exists $i \in V$ such that

$$v(x_i, q) \cdot \sum_{\{j:ij\in E\}} w_{ij} v(x_i, A_{ij}x_j) = v(x_i, q) \cdot r_i > 0,$$

and thus

$$v_i \cdot P_{x_i}(r_i) = -\frac{1}{d(q, x_i)} P_{x_i}(v(x_i, q)) \cdot P_{x_i}(r_i) = -\frac{1}{d(q, x_i)} v(x_i, q) \cdot r_i < 0. \qquad \square$$

In the rest of this subsection we will prove Lemma 2.5 by contradiction. Let us first sketch the idea of the proof. Assume $\sum_{i\in V} d(x_i, q)^2$ is very large. Then by a standard compactness argument there exists a long edge ij in the geodesic mapping $\varphi[x_1, \dots, x_n]$. Assume $d(q, x_i) \geq d(q, x_j)$. Then the corresponding long edge $\gamma_{x_i, A_{ij}x_j}$ in $\widetilde{M}$ is pulling x_i towards q. This implies that there exists another long edge $\gamma_{x_i, A_{ik}x_k}$ dragging x_i away from q, otherwise the residue vector r_i would not drag x_i away from q. It can be shown that $d(q, x_k) > d(q, x_i)$. Repeating the above steps, we can find an arbitrarily long sequence of vertices such that the distance from each of these vertices to q is increasing. This is impossible as we only have finitely many vertices.

Here is a list of useful properties, where (a), (e), (f), (g) and (h) serve directly as building blocks of the proof of Lemma 2.5, (b) and (d) are used to prove (e), and (c) is used to prove (h). The three triangles in Figure 2, from left to right, illustrate the geodesic triangles appearing in (b), (c) and (d) of Lemma 2.6, respectively.

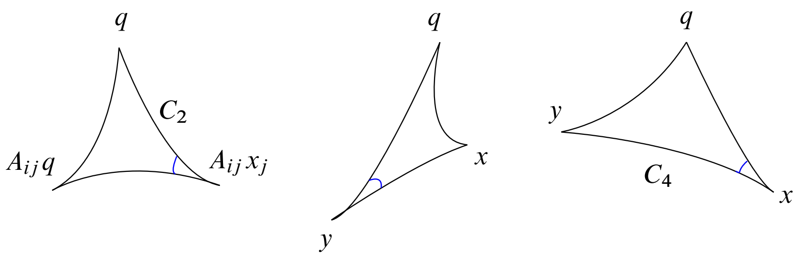

Figure 2: Triangles in (b), (c) and (d).

Lemma 2.6 (a) *For any constant $C>0$ there is a constant $C_1=C_1(M,T,\psi,C)>0$ such that, if*

$$\sum_{i\in V} d(x_i,q)^2 \geq C_1,$$

then

$$\max_{ij\in E} d(x_i,A_{ij}x_j)\geq C.$$

(b) *There exists a constant $C_2=C_2(M,T,\psi)>0$ such that, if*

$$d(A_{ij}x_j,q)\geq C_2,$$

then

$$\angle(A_{ji}q)x_jq=\angle q(A_{ij}x_j)(A_{ij}q)\leq \tfrac{1}{8}\pi.$$

(c) *There exists a constant $C_3>0$ such that, if $x,y\in\widetilde{M}$ satisfy*

$$d(y,q)\geq d(x,q)+C_3,$$

then

$$\angle xyq\leq \tfrac{1}{4}\pi.$$

(d) *There exists a constant $C_4>0$ such that, if $x,y\in\widetilde{M}$ satisfy*

$$d(x,y)\geq C_4 \quad \text{and} \quad d(x,q)\geq d(y,q),$$

then

$$\angle yxq\leq \tfrac{1}{8}\pi.$$

(e) *For any constant $C>0$ there is a constant $C_5=C_5(M,T,\psi,C)>0$ such that, if*

$$\max_{ij\in E} d(x_i,A_{ij}x_j)\geq C_5,$$

then there exists $ij\in E$ such that

$$\frac{v(x_i,q)}{|v(x_i,q)|}\cdot v(x_i,A_{ij}x_j)\geq C.$$

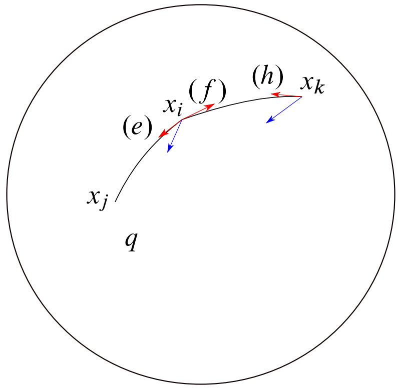

Figure 3: Vertices leaving the point q.

(f) *For any constant* $C>0$ *there is a constant* $C_6=C_6(M,T,\psi,\lambda_w,C)>0$ *such that, if*

$$\frac{v(x_i,q)}{|v(x_i,q)|}\cdot v(x_i,A_{ij}x_j)\geq C_6$$

for some edge $ij\in E$, *then there exists* $ik\in E$ *such that*

$$\frac{v(x_i,q)}{|v(x_i,q)|}\cdot v(x_i,A_{ik}x_k)\leq -C.$$

(g) *For any constant* $C>0$ *there is a constant* $C_7=C_7(M,T,\psi,C)>0$ *such that, if*

$$\frac{v(x_i,q)}{|v(x_i,q)|}\cdot v(x_i,A_{ik}x_k)\leq -C_7,$$

then

$$d(x_k,q)\geq d(x_i,q)+C.$$

(h) *For any constant* $C>0$ *there is a constant* $C_8=C_8(M,T,\psi,C)>0$ *such that, if*

$$d(x_j,q)\geq d(x_i,q)+C_8,$$

then

$$\frac{v(x_j,q)}{|v(x_j,q)|}\cdot v(x_j,A_{ji}x_i)\geq C.$$

Proof of Lemma 2.5 assuming Lemma 2.6 For any $C>0$ there exists a sufficiently large constant $\widetilde{C}=\widetilde{C}(M,T,\psi,\lambda_w,C)$ determined by (a), (e), (f) and (g) in Lemma 2.6 such that, if

$$\sum_{i\in V} d(x_i,q)^2\geq \widetilde{C},$$

then there exist three vertices x_i, x_j and x_k, shown in Figure 3, with

$$d(x_k,q)\geq d(x_j,q)+C.$$

Moreover, by (h), (f) and (g) of Lemma 2.6, we can find another vertex x_l such that

$$d(x_l, q) \geq d(x_k, q) + C \geq d(x_j, q) + 2C$$

if the constant $\widetilde{C}(M, T, \psi, \lambda_w, C)$ is sufficiently large.

Inductively, we can find a sequence $i_1, \dots, i_{n+1} \in V$ such that

$$d(x_{i_1}, q) > d(x_{i_2}, q) > \dots > d(x_{i_{n+1}}, q).$$

This contradicts the fact that V only has n different elements. □

Proof of Lemma 2.6 (a) By a standard compactness argument, the set

$$\{\varphi \in \widetilde{X} : \max_{ij \in E} \operatorname{length}(\varphi_{ij}([0,1])) \leq C\}$$

is a compact subset of $\widetilde{X}$. Notice that $(x_1, \dots, x_n) \mapsto \varphi[x_1, \dots, x_n]$ is a homeomorphism from $\mathcal{M}$ to $\widetilde{X}$, and

$$\operatorname{length}(\varphi_{ij}([0,1])) = d(x_i, A_{ij} x_j).$$

Therefore

$$\Bigl\{(x_1, \dots, x_n) \in \mathcal{M} : \max_{ij \in E} \; d(x_i, A_{ij} x_j) \leq C\Bigr\}$$

is compact, and the conclusion follows.

(b) We claim that the constant C_2 determined by

$$\sinh C_2 = \frac{\max_{ij \in E} \sinh d(A_{ij} q, q)}{\sin \frac{1}{8}\pi}$$

is satisfactory. Let $\triangle ABC$ be the hyperbolic triangle with corresponding edge lengths

$$a = d(A_{ij} x_j, q), \quad b = d(A_{ij} x_j, A_{ij} q) \quad \text{and} \quad c = d(A_{ij} q, q).$$

Since $\widetilde{M}$ is a CAT(-1) space, it suffices to show that $\angle C \leq \frac{1}{8}\pi$. By the hyperbolic law of sine,

$$\sin \angle C = \frac{\sinh c \cdot \sin \angle A}{\sinh a} \leq \frac{\max_{ij \in E} \sinh d(A_{ij} q, q) \cdot 1}{\sinh C_2} = \sin \tfrac{1}{8}\pi.$$

(c) We claim that the constant C_3 determined by

$$\sinh C_3 = \frac{1}{\sin \frac{1}{8}\pi}$$

is satisfactory. Let $\triangle ABC$ be the hyperbolic triangle with corresponding edge lengths

$$a = d(x, y), \quad b = d(y, q) \quad \text{and} \quad c = d(x, q).$$

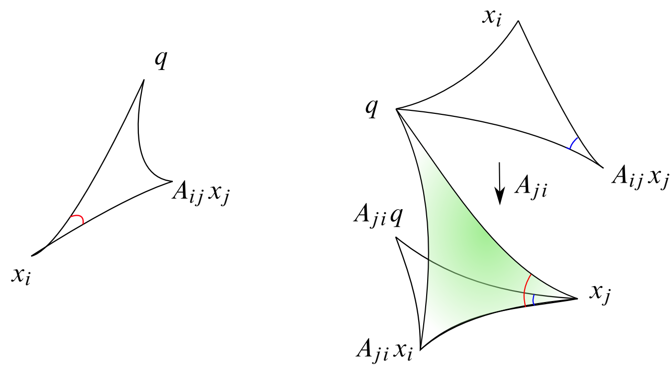

Figure 4: Triangles in (e).

Since $\widetilde{M}$ is a CAT(-1) space it suffices to show that $\angle C \leq \frac{1}{8}\pi$. By the hyperbolic law of sine,

$$\sin\angle C = \frac{\sinh c\cdot\sin\angle B}{\sinh b} \leq \frac{\sinh c}{\sinh b} \leq \frac{\sinh c}{\sinh(c+C_3)} \leq \frac{\sinh c}{\sinh c\cdot\sinh C_3} = \sin\tfrac{1}{8}\pi.$$

(d) We claim that the constant C_4 determined by

$$\sin^2\tfrac{1}{8}\pi\cdot\cosh C_4 = 2$$

is satisfactory. Let $\triangle ABC$ be the hyperbolic triangle with corresponding edge lengths

$$a = d(x,y), \quad b = d(y,q) \quad \text{and} \quad c = d(x,q).$$

Since $\widetilde{M}$ is a CAT(-1) space it suffices to show that $\angle B \leq \frac{1}{8}\pi$. By the hyperbolic law of cosine,

$$\cos A = -\cos B\cos C + \sin B\sin C\cosh a.$$

Then

$$2 \geq \sin B\sin C\cosh a \geq \sin B\sin C\cosh C_4 = 2\cdot\frac{\sin B\sin C}{\sin^2\frac{1}{8}\pi} \geq 2\cdot\frac{\sin^2 B}{\sin^2\frac{1}{8}\pi}.$$

Thus, $\angle B \leq \frac{1}{8}\pi$.

(e) We claim that the constant C_5 determined by

$$C_5 = \max\{C_4, 2C_2, \sqrt{2}C\}$$

is satisfactory. Assume $ij \in E$ and $d(x_i, A_{ij}x_j) \geq C_5$. Then we have the two cases shown in Figure 4.

If $d(x_i,q) \geq d(A_{ij}x_j,q)$, then, by (d),

$$\angle(A_{ij}x_j)x_i q \leq \tfrac{1}{8}\pi \leq \tfrac{1}{4}\pi,$$

and

$$\frac{v(x_i,q)}{|v(x_i,q)|}\cdot v(x_i,A_{ij}x_j)=\cos(\angle(A_{ij}x_j)x_iq)\cdot d(x_i,A_{ij}x_j)\geq \frac{1}{\sqrt{2}}C_5\geq C.$$

If $d(x_i,q)\leq d(A_{ij}x_j,q)$, then $d(A_{ij}x_j,q)\geq C_2$. By (b) and (d),

$$\angle(A_{ji}q)x_jq\leq \tfrac{1}{8}\pi \quad\text{and}\quad \angle(A_{ji}q)x_j(A_{ji}x_i)=\angle q(A_{ij}x_j)x_i\leq \tfrac{1}{8}\pi,$$

and $\angle qx_j(A_{ji}x_i)\leq \frac{1}{4}\pi$. Therefore,

$$\frac{v(x_j,q)}{|v(x_j,q)|}\cdot v(x_j,A_{ji}x_i)=\cos(\angle(A_{ji}x_j)x_jq)\cdot d(x_j,A_{ji}x_i)\geq \frac{1}{\sqrt{2}}C_5\geq C.$$

(f) We claim that the constant C_6 determined by

$$C_6=n\lambda_w\cdot C$$

is satisfactory. If not, for any $ik\in E$,

$$\frac{v(x_i,q)}{|v(x_i,q)|}\cdot v(x_i,A_{ik}x_k)>-C.$$

Then

$$\begin{aligned}0\geq \frac{v(x_i,q)}{|v(x_i,q)|}\cdot\sum_{ik\in E}w_{ik}v(x_i,A_{ik}x_k)&>w_{ij}C_6+\sum_{ik\in E}w_{ik}(-C)\\ &\geq w_{ij}C_6+\sum_{ik\in E}\lambda_w w_{ij}(-C)\geq w_{ij}(C_6-n\lambda_wC)\geq 0,\end{aligned}$$

which is a contradiction.

(g) We claim that $C_7=C+\max_{ij\in E}d(A_{ij}q,q)$ is satisfactory. Notice that

$$d(A_{ij}x_j,q)=d(x_j,A_{ji}q)\leq d(x_j,q)+d(q,A_{ji}q)\leq d(x_j,q)+\max_{ij\in E}d(A_{ij}q,q).$$

By Lemma 2.2(i),

$$\begin{aligned}d(A_{ij}x_j,q)\geq |v(x_i,A_{ij}x_j)-v(x_i,q)|&\geq -(v(x_i,A_{ij}x_j)-v(x_i,q))\cdot\frac{v(x_i,q)}{|v(x_i,q)|}\\ &=C_7+|v(x_i,q)|=C_7+d(x_i,q).\end{aligned}$$

Then

$$d(x_j,q)-d(x_i,q)\geq C_7-\max_{ij\in E}d(A_{ij}q,q)=C.$$

(h) We claim that the constant C_8 determined by

$$C_8=\max\{C_3,\sqrt{2}C\}+\max_{ij\in E}d(A_{ij}q,q)$$

is satisfactory. Notice that

$$\begin{aligned} d(x_j, q) &\geq d(x_i, q) + C_8 \geq d(x_i, A_{ij} q) - d(A_{ij} q, q) + C_8 \\ &\geq d(A_{ji} x_i, q) + \max\{C_3, \sqrt{2} C\}. \end{aligned}$$

Then, by (c), $\angle(A_{ji} x_i) x_j q \leq \frac{1}{4}\pi$, and by the triangle inequality,

$$d(x_j, A_{ji} x_i) \geq d(x_j, q) - d(A_{ji} x_i, q) \geq \sqrt{2} C.$$

Therefore,

$$\frac{v(x_j, q)}{|v(x_j, q)|} \cdot v(x_j, A_{ji} x_i) = \cos(\angle(A_{ji} x_i) x_j q) \cdot d(x_j, A_{ji} x_i) \geq \frac{1}{\sqrt{2}} \cdot \sqrt{2} C = C. \quad \square$$

2.3 Continuity

Proof of continuity in Theorem 1.5 If Φ is not continuous, there exists $\epsilon > 0$, a weight w and a sequence of weights $w^{(k)}$ such that

(i) the $w^{(k)}$ converge to w, and

(ii) $d_{\widetilde{X}}(\Phi(w^{(k)}), \Phi(w)) \geq \epsilon$ for any $k \geq 1$.

By the stronger existence result Theorem 2.3, the sequence $\Phi(w^{(k)})$ is in some fixed compact subset K' of $\widetilde{X}$. By picking a subsequence, we may assume that $\Phi(w^{(k)})$ converges to some $\varphi \in \widetilde{X}$. Since $\Phi(w^{(k)})$ is $w^{(k)}$–balanced, then, by the continuity of the residue vectors r_i, φ is w–balanced and thus $\Phi(w) = \varphi$, which contradicts that $\Phi(w^{(k)})$ does not converge to $\Phi(w)$. $\square$

3 Proof of Theorem 1.6

3.1 Setup and preparation

Assume $\varphi \in \widetilde{X}$ is w–balanced for some weight w. We will prove that φ is an embedding. Recall that $q_i = \varphi(i)$ for each $i \in V$, and denote by l_{ij} the length of $\varphi_{ij}([0,1])$ for any $ij \in E$. It is not difficult to show that φ has a continuous extension $\tilde{\varphi}$ defined on $|T|$ such that for any triangle $\sigma \in F$ a continuous lifting map Φ_σ of $\tilde{\varphi}|_\sigma$ from σ to $\widetilde{M}$ will map σ to

(i) a geodesic triangle in $\widetilde{M}$ homeomorphically if $\varphi(\partial\sigma)$ does not degenerate to a geodesic, and

(ii) $\Phi_\sigma(\partial\sigma)$ if $\varphi(\partial\sigma)$ degenerates to a geodesic.

The main tool we use to prove Theorem 1.6 is the Gauss–Bonnet formula. We will need to define the inner angles for each triangle in $\varphi(T^{(1)})$, even for the degenerate triangles.

A convenient way is to assign a "direction" to each edge, even for the degenerate edges with zero length.

Definition 3.1 A *direction field* is a map $v\colon \vec{E} \to TM$ satisfying

(i) $v_{ij} \in T_{q_i}M$ for any $(i,j) \in \vec{E}$, and

(ii) $|v_{ij}| = 1$ for any $(i,j) \in \vec{E}$.

Given a direction field v, define the inner angle of the triangle $\sigma = \triangle ijk$ at the vertex i as

$$\theta^i_\sigma = \theta^i_\sigma(v) = \angle v_{ij} 0 v_{ik} = \arccos(v_{ij} \cdot v_{ik}),$$

where 0 is the origin and $\angle v_{ij} 0 v_{ik}$ is the angle between v_{ij} and v_{ik} in $T_{q_i}M$.

A direction field v assigns a unit tangent vector in $T_{q_i}M$ to each directed edge starting from i, even if their lengths are zero. It determines the inner angles in T.

Definition 3.2 A direction field v is *admissible* if

(i) $v_{ij} = \varphi'_{ij}(0)/l_{ij}$ if $l_{ij} > 0$,

(ii) $v_{ij} = -v_{ji}$ in $T_{q_i}M = T_{q_j}M$ if $l_{ij} = 0$,

(iii) for a fixed vertex $i \in V$, if $l_{ij} = 0$ for every neighbor j of i, then there exist neighbors k and k' of i such that $v_{ik} = -v_{ik'}$, and

(iv) if $\sigma = \triangle ijk \in F$ and $l_{ij} = l_{jk} = l_{ik} = 0$, then $\theta^i_\sigma(v) + \theta^j_\sigma(v) + \theta^k_\sigma(v) = \pi$.

An admissible direction field encodes the directions of the nondegenerate edges in $\varphi(T^{(1)})$, and the induced angle sum of a degenerate triangle is always π. Then, for any admissible v and triangle $\sigma \in F$, by the Gauss–Bonnet formula,

$$\pi = \sum_{i\in\sigma} \theta^i_\sigma(v) - \int_{\Phi_\sigma(\sigma)} K\,dA \geq \sum_{i\in\sigma} \theta^i_\sigma(v) - \int_{\tilde{\varphi}(\sigma)} K\,d\tilde{A}. \tag{4}$$

Here dA (resp. $d\tilde{A}$) is the area form on (M,g) (resp. $(\widetilde{M},\tilde{g})$).

The concept of the direction field is similar to the *discrete one-form* defined in [16].

3.2 Proof of Theorem 1.6

The proof of Theorem 1.6 uses the four lemmas below. We will postpone their proofs to the subsequent subsections.

Lemma 3.3 *If v is admissible and $\theta = \theta(v)$, then, for any $i \in V$,*

$$\sum_{\{\sigma : i \in \sigma\}} \theta_\sigma^i = 2\pi,$$

and $\tilde{\varphi}(\sigma) \cap \tilde{\varphi}(\sigma')$ has area 0 for any $\sigma, \sigma' \in F$.

Based on Lemma 3.3, if admissible direction fields exist, the image of the star of each vertex determined by $\tilde{\varphi}$ does not contain any flipped triangles overlapping with each other. If $\tilde{\varphi}(\sigma)$ does not degenerate to a geodesic arc for any triangle $\sigma \in F$, then $\tilde{\varphi}$ is locally homeomorphic and thus globally homeomorphic as a degree-one map. Therefore, we only need to exclude the existence of degenerate triangles.

Define an equivalence relation on V as follows. Two vertices i and j are equivalent if there exists a sequence of vertices $i = i_0, i_1, \dots, i_k = j$ such that $l_{i_0 i_1} = \cdots = l_{i_{k-1} i_k} = 0$. This equivalence relation introduces a partition $V = V_1 \cup \cdots \cup V_m$. Denote by $y_k \in M$ the only point in $\varphi(V_k)$. For any $x \in M$ and $u, v \in T_x M$, write $u \parallel v$ if u and v are parallel, ie there exists $(\alpha, \beta) \neq (0,0)$ such that $\alpha u + \beta v = 0$.

There are plenty of choices of admissible direction fields:

Lemma 3.4 *For any $v_1 \in T_{y_1} M, \dots, v_m \in T_{y_m} M$, there exists an admissible v such that $v_{ij} \parallel v_k$ if $i \in V_k$ and $l_{ij} = 0$.*

For any V_k with at least two vertices, the image of its "neighborhood" lies in a geodesic:

Lemma 3.5 *If $|V_k| \geq 2$, then there exists $v_k \in T_{y_k} M$ such that $v_k \parallel \varphi'_{ij}(0)$ if $i \in V_k$ and $l_{ij} > 0$.*

Now let v_k be as in Lemma 3.5 if $|V_k| \geq 2$, and arbitrary if $|V_k| = 1$. Then construct an admissible direction field v as in Lemma 3.4, with induced inner angles $\theta_\sigma^i = \theta_\sigma^i(v)$. If the image of a triangle σ under φ degenerates to a geodesic, then its inner angles θ_σ^i are π or 0. Let $F' \neq \varnothing$ be the set of degenerate triangles under φ.

Lemma 3.6 *If $\sigma \in F'$, $i \in \sigma$ and $\theta_\sigma^i = \pi$, then $\sigma' \in F'$ for any σ' in the star neighborhood of the vertex i.*

Let Ω be a connected component of the interior of $\bigcup \{\sigma : \sigma \in F'\} \subset |T|$, and $\widetilde{\Omega}$ be the completion of Ω under the natural path metric on Ω. Notice that $\widetilde{\Omega}$ could be different from the closure of Ω in M.

Since $\tilde{\varphi}$ is surjective $F' \neq F$, $\Omega \neq |T|$ and $\widetilde{\Omega}$ has nonempty boundary. Then $\widetilde{\Omega}$ is a connected surface with a natural triangulation $T' = (V', E', F')$, and

$$\chi(\widetilde{\Omega}) = 2 - 2 \times (\text{genus of } \widetilde{\Omega}) - \#\{\text{boundary components of } \widetilde{\Omega}\} \leq 1.$$

Assume V'_I is the set of interior vertices, V'_B is the set of boundary vertices, E'_I is the set of interior edges and E'_B is the set of boundary edges of $\widetilde{\Omega}$. Then $|V'_B| = |E'_B|$ and, by Lemma 3.6, if $i \in V'_B$ and $i \in \sigma$ then $\theta^i_\sigma = 0$. Therefore,

$$\pi|F'| = \sum_{\substack{\sigma \in F' \\ i \in \sigma}} \theta^i_\sigma = \sum_{i \in V'_I} \sum_{\{\sigma \in F' : i \in \sigma\}} \theta^i_\sigma = 2\pi |V'_I|.$$

Thus,

$$\begin{aligned} 1 \geq \chi(\widetilde{\Omega}) &= |V'| - |E'| + |F'| = |V'_I| + |V'_B| - |E'_I| - |E'_B| + |F'| \\ &= |V'_B| - |E'_I| - |E'_B| + \tfrac{3}{2}|F'| = -|E'_I| + \tfrac{3}{2}|F'| \\ &= -|E'_I| + \tfrac{1}{2}(|E'_B| + 2|E'_I|) = \tfrac{1}{2}|E'_B|. \end{aligned}$$

Therefore, $|V'_B| = |E'_B| \leq 2$. Since $\widetilde{\Omega}$ has nonempty boundary, $|E'_B| = 1$ or 2. In either case, it contradicts the fact that T is a simplicial complex. The proof of Theorem 1.6 is now completed.

3.3 Proof of Lemma 3.3

We claim that, for any $i \in V$,

$$\sum_{\{\sigma : i \in \sigma\}} \theta^i_\sigma \geq 2\pi.$$

If $l_{ij} = 0$ for any neighbor j of i, this is a consequence of Definition 3.2(iii). If $l_{ij} \neq 0$, by the w–balanced condition, $\{\varphi'_{ij}(0)/l_{ij} : ij \in E\}$ should not be contained in any open half unit circle, and the angle sum around i should be at least 2π.

By the fact that $\tilde{\varphi}$ is surjective and (4),

$$\sum_{i \in V}\Bigl(2\pi - \sum_{\{\sigma : i \in \sigma\}} \theta^i_\sigma\Bigr) + \sum_{\sigma \in F} \int_{\tilde{\varphi}(\sigma)} K\, dA \leq \sum_{\sigma \in F} \int_{\tilde{\varphi}(\sigma)} K\, dA \leq \int_M K\, dA = 2\pi\chi(M)$$

and

$$\begin{aligned} \sum_{i \in V}\Bigl(2\pi - \sum_{\{\sigma : i \in \sigma\}} \theta^i_\sigma\Bigr) + \sum_{\sigma \in F} \int_{\tilde{\varphi}(\sigma)} K\, dA &\geq 2\pi|V| - \sum_{\sigma \in F}\Bigl(\sum_{i \in \sigma} \theta^i_\sigma - \int_{\tilde{\varphi}(\sigma)} K\, dA\Bigr) \\ &= 2\pi\chi(M). \end{aligned}$$

Hence, the inequalities above are equalities. This fact implies that

$$\sum_{i\in V}\Bigl(2\pi-\sum_{\{\sigma:i\in\sigma\}}\theta^i_\sigma\Bigr)=0.$$

Since each term in this summation is nonpositive, $\sum_{\{\sigma:i\in\sigma\}}\theta^i_\sigma=2\pi$. The statement on the area follows similarly.

3.4 Proof of Lemma 3.4

We claim that, for any k, there exists a map $h\colon V_k\to\mathbb{R}$ such that

(i) $h(i)\neq h(j)$ if $i\neq j$, and

(ii) for a fixed $i\in V_k$, if $l_{ij}=0$ for any neighbor j of i, then there exist neighbors j and j' of i in V_k such that $h(j)<h(i)<h(j')$.

Given such h, set v as

$$v_{ij}=\begin{cases}\operatorname{sgn}[h(j)-h(i)]\cdot v_k & \text{if } i\in V_k \text{ and } l_{ij}=0,\\ \varphi'_{ij}(0) & \text{if } l_{ij}>0,\end{cases}$$

where sgn is the sign function. It is easy to verify that v is satisfactory.

To construct h, we prove a more general statement:

Lemma 3.7 *Assume $G=(V',E')$ is a subgraph of the 1–skeleton $T^{(1)}$, and $E'\neq E$. Define*

$$\operatorname{int}(G)=\{i\in V':ij\in E\Rightarrow ij\in E'\},$$

and $\partial G=V'-\operatorname{int}(G)$. Then there exists $h\colon V'\to\mathbb{R}$ such that

(i) *$h(i)\neq h(j)$ if $i\neq j$, and*

(ii) *any $i\in\operatorname{int}(G)$ has neighbors j and j' in V' such that $h(j)<h(i)<h(j')$.*

Proof We proceed by induction on the size of V'. The case $|V'|=1$ is trivial. For the case $|V'|\geq 2$, first notice that $|\partial G|\geq 2$ for any proper subgraph G of $T^{(1)}$. Assign distinct values $\bar h(i)$ to each $i\in\partial G$, then solve the discrete harmonic equation

$$\sum_{\{j:ij\in E\}}(\bar h(j)-\bar h(i))=0\quad\text{for all } i\in\operatorname{int}(G)$$

with the given Dirichlet boundary condition on ∂G.

Let $s_1 < \cdots < s_k$ be the distinct values that appear in $\{\bar{h}(i) : i \in V'\}$. Then consider the subgraphs $G_i = (V_i', E_i')$ defined by

$$V_i' = \{j \in V' : h(j) = s_i\} \quad \text{and} \quad E_i' = \{jj' \in E' : j, j' \in V_i'\}.$$

Notice that $|\partial G| \geq 2$, so $k \geq 2$ and $|V_i'| < |V'|$ for any $i = 1, \dots, k$. By the induction hypothesis, there exists a function $h_i : V_i' \to \mathbb{R}$ such that

(i) $h_i(j) \neq h_i(j')$ if $j \neq j'$, and

(ii) any $j \in \operatorname{int}(G_i)$ has neighbors j' and j'' in V_i' such that $h_i(j') < h_i(j) < h_i(j'')$.

Define $h_i(j) = 0$ if $j \notin V_i'$. Then, for sufficiently small positive $\epsilon_1, \dots, \epsilon_k$,

$$h = \bar{h} + \sum_{i=1}^{k} \epsilon_i h_i$$

is the desired function. □

3.5 Proof of Lemma 3.5

We must prove that if $i, i' \in V_k$, $ij, i'j' \in E$, $l_{ij} > 0$ and $l_{i'j'} > 0$, then $\varphi_{ij}'(0) \parallel \varphi_{i'j'}'(0)$. Let

$$D = \Big(\bigcup_{i,i',i'' \in V_k} \triangle i i' i''\Big) \cup \Big(\bigcup_{i,i' \in V_k} i i'\Big),$$

which is a closed path-connected set in $|T|$. For any $i \in V_k$, we have $i \in \partial D$ if and only if there exists $ij \in E$ with $l_{ij} > 0$. Therefore, it suffices to prove that

(i) $\varphi_{ij}'(0) \parallel \varphi_{ij'}'(0)$ for any $i \in V_k$ and edges ij and ij' with $l_{ij} > 0$ and $l_{ij'} > 0$,

(ii) for any $ij \in E$ satisfying $ij \subset \partial D$, there exists $m \in V - V_k$ such that $\triangle ijm \in F$, and thus $\varphi_{im}'(0) = \varphi_{jm}'(0)$, and

(iii) ∂D is connected.

For part (i), if it is not true then there exists $i \in V_k$, $ij \in E$ and $ij' \in E$ such that $l_{ij} > 0$, $l_{ij'} > 0$ and $\varphi_{ij}'(0)$ is not parallel to $\varphi_{ij'}'(0)$. Assuming this claim, by the w–balanced condition, there exists $ij'' \in E$ with $l_{ij''} > 0$, and the three vectors $\varphi_{ij}'(0)$, $\varphi_{ij'}'(0)$ and $\varphi_{ij''}'(0)$ are not contained in any closed half-space in $T_{q_i} M$. Assume $im \in E$, $l_{im} = 0$ and, without loss of generality, ij, im, ij' and ij'' are ordered counterclockwise in the one-ring neighborhood of i in T. By Lemma 3.4, there exists an admissible v such that $v_{im} \parallel \varphi_{ij''}'(0)$. By possibly changing a sign, we may assume that $v_{im} = \varphi_{ij''}'(0)/l_{ij''}$.

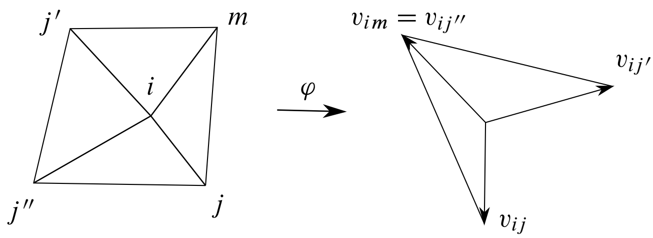

Figure 5: Overlapping triangles lead to angle surplus.

Then, as Figure 5 shows, a contradiction follows:

$$\begin{aligned} 2\pi &= \sum_{\sigma i \in \sigma} \theta^i_\sigma \geq \angle v_{ij} 0 v_{im} + \angle v_{im} 0 v_{ij'} + \angle v_{ij'} 0 v_{ij''} + \angle v_{ij''} 0 v_{ij} \\ &= 2\angle v_{ij} 0 v_{ij''} + 2\angle v_{ij''} 0 v_{ij'} > 2\pi. \end{aligned}$$

Part (ii) is straightforward, so we will prove part (iii). By our assumption on the extension $\tilde{\varphi}$, $\tilde{\varphi}(D)$ contains only one point. Then the embedding map $i_D = \psi^{-1} \circ (\psi|_D)$ from D to $|T|$ is homotopic to the constant map $\psi^{-1} \circ (\tilde{\varphi}|_D)$, meaning that D is contractible in $|T|$. If ∂D contains at least two boundary components, then it is not difficult to show that $|T| - D$ has a connected component D' homeomorphic to an open disk. Let $\Phi_D : D \to \widetilde{M}$ be a lifting of $\tilde{\varphi}|_D$. Then $\Phi_D(\partial D') \subset \Phi_D(D)$ contains only a single point $x \in \widetilde{M}$. So, by the w–balanced condition, it is not difficult to derive a maximum principle and show that $\tilde{\varphi}|_{D'}$ equals the constant x. Then, by the definition of D, it is easy to see that D' should be a subset of D, which is a contradiction.

3.6 Proof of Lemma 3.6

Assume ij and ij' are two edges in σ. If the conclusion is not true, then there exists $ik \in E$ such that $l_{ik} > 0$ and v_{ik} is not parallel to v_{ij}. Notice that $v_{ij} = -v_{ij'}$, and

$$2\pi = \sum_{\{\sigma \in F : i \in \sigma\}} \theta^i_\sigma \geq \angle v_{ij} 0 v_{ij'} + \angle v_{ij} 0 v_{ik} + \angle v_{ij'} 0 v_{ik} = 2\pi.$$

Thus, equality holds in the above inequality, and for any $ik' \in E$, $v_{ik'}$ should be on the half circle that contains v_{ij}, v_{ik} and $v_{ij'}$. If v_m is the midpoint of this half circle, then

$$v_m \cdot \sum_{\{j : ij \in E\}} w_{ij} l_{ij} v_{ij} \geq w_{ik} l_{ik} v_m \cdot v_{ik} > 0.$$

This contradicts the fact that φ is w–balanced.

References

[1] **M Awartani**, **D W Henderson**, *Spaces of geodesic triangulations of the sphere*, Trans. Amer. Math. Soc. 304 (1987) 721–732 MR Zbl

[2] **E D Bloch**, **R Connelly**, **D W Henderson**, *The space of simplexwise linear homeomorphisms of a convex* 2*–disk*, Topology 23 (1984) 161–175 MR Zbl

[3] **M R Bridson**, **A Haefliger**, *Metric spaces of non-positive curvature*, Grundl. Math. Wissen. 319, Springer (1999) MR Zbl

[4] **D Burago**, **Y Burago**, **S Ivanov**, *A course in metric geometry*, Graduate Studies in Math. 33, Amer. Math. Soc., Providence, RI (2001) MR Zbl

[5] **P Buser**, **H Karcher**, *Gromov's almost flat manifolds*, Astérisque 81, Soc. Math. France, Paris (1981) MR Zbl

[6] **S S Cairns**, *Isotopic deformations of geodesic complexes on the* 2*–sphere and on the plane*, Ann. of Math. 45 (1944) 207–217 MR Zbl

[7] **E W Chambers**, **J Erickson**, **P Lin**, **S Parsa**, *How to morph graphs on the torus*, from "Proceedings of the 2021 ACM–SIAM symposium on discrete algorithms" (D Marx, editor), SIAM, Philadelphia, PA (2021) 2759–2778 MR

[8] **Y Colin de Verdière**, *Comment rendre géodésique une triangulation d'une surface?*, Enseign. Math. 37 (1991) 201–212 MR Zbl

[9] **R Connelly**, **D W Henderson**, **C W Ho**, **M Starbird**, *On the problems related to linear homeomorphisms*, *embeddings*, *and isotopies*, from "Continua, decompositions, manifolds" (R H Bing, W T Eaton, M P Starbird, editors), Univ. Texas Press, Austin, TX (1983) 229–239 MR

[10] **C J Earle**, **J Eells**, *A fibre bundle description of Teichmüller theory*, J. Differential Geometry 3 (1969) 19–43 MR Zbl

[11] **J Erickson**, **P Lin**, *Planar and toroidal morphs made easier*, from "Graph drawing and network visualization" (H C Purchase, I Rutter, editors), Lecture Notes in Comput. Sci. 12868, Springer (2021) 123–137 MR Zbl

[12] **M S Floater**, *Mean value coordinates*, Comput. Aided Geom. Design 20 (2003) 19–27 MR Zbl

[13] **M S Floater**, *One-to-one piecewise linear mappings over triangulations*, Math. Comp. 72 (2003) 685–696 MR Zbl

[14] **M S Floater**, **C Gotsman**, *How to morph tilings injectively*, J. Comput. Appl. Math. 101 (1999) 117–129 MR Zbl

[15] **J Gaster**, **B Loustau**, **L Monsaingeon**, *Computing discrete equivariant harmonic maps*, preprint (2018) arXiv 1810.11932

[16] **S J Gortler**, **C Gotsman**, **D Thurston**, *Discrete one-forms on meshes and applications to 3D mesh parameterization*, Comput. Aided Geom. Design 23 (2006) 83–112 MR Zbl

[17] **K Grove**, **H Karcher**, *How to conjugate C^1–close group actions*, Math. Z. 132 (1973) 11–20 MR Zbl

[18] **J Hass**, **P Scott**, *Simplicial energy and simplicial harmonic maps*, Asian J. Math. 19 (2015) 593–636 MR Zbl

[19] **H Karcher**, *Riemannian center of mass and mollifier smoothing*, Comm. Pure Appl. Math. 30 (1977) 509–541 MR Zbl

[20] **H Karcher**, *Riemannian center of mass and so called karcher mean*, preprint (2014) arXiv 1407.2087

[21] **Y Luo**, *Spaces of geodesic triangulations of surfaces*, Discrete Comput. Geom. 68 (2022) 709–727 MR Zbl

[22] **Y Luo**, **T Wu**, **X Zhu**, *The deformation spaces of geodesic triangulations of flat tori*, preprint (2021) arXiv 2107.05159

[23] **S Smale**, *Diffeomorphisms of the 2–sphere*, Proc. Amer. Math. Soc. 10 (1959) 621–626 MR Zbl

[24] **W T Tutte**, *How to draw a graph*, Proc. London Math. Soc. 13 (1963) 743–767 MR Zbl

Department of Mathematics, Rutgers University
New Brunswick, NJ, United States

Center of Mathematical Sciences and Applications, Harvard University
Cambridge, MA, United States

Department of Mathematics, Rutgers University
New Brunswick, NJ, United States

yl1594@rutgers.edu, mike890505@gmail.com, xiaoping.zhu@rutgers.edu

Proposed: Benson Farb Received: 18 July 2021
Seconded: David Fisher, Mladen Bestvina Revised: 16 January 2022

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

Volume 27 Issue 8 (pages 2937–3385) 2023