Pacific Journal of Mathematics

Volume 293 No. 1 March 2018

PACIFIC JOURNAL OF MATHEMATICS

Founded in 1951 by E. F. Beckenbach (1906–1982) and F. Wolf (1904–1989)

msp.org/pjm

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

The subscription price for 2018 is US $475/year for the electronic version, and $640/year for print and electronic.
Subscriptions, requests for back issues and changes of subscriber address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163, U.S.A. The Pacific Journal of Mathematics is indexed by Mathematical Reviews, Zentralblatt MATH, PASCAL CNRS Index, Referativnyi Zhurnal, Current Mathematical Publications and Web of Knowledge (Science Citation Index).

The Pacific Journal of Mathematics (ISSN 0030-8730) at the University of California, c/o Department of Mathematics, 798 Evans Hall #3840, Berkeley, CA 94720-3840, is published twelve times a year. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices. POSTMASTER: send address changes to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163.

PJM peer review and production are managed by EditFlow® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.1

LARGE-SCALE RIGIDITY PROPERTIES OF THE MAPPING CLASS GROUPS

BRIAN H. BOWDITCH

We study the coarse geometry of the mapping class group of a compact orientable surface. We show that, apart from a few low-complexity cases, any quasi-isometric embedding of a mapping class group into itself agrees up to bounded distance with a left multiplication. In particular, such a map is a quasi-isometry. This is a strengthening of the result of Hamenstädt and of Behrstock, Kleiner, Minsky and Mosher that the mapping class groups are quasi-isometrically rigid. In the course of proving this, we also develop the general theory of coarse median spaces and median metric spaces with a view to applications to Teichmüller space, and related spaces.

1. Introduction

One of the main aims of this paper is to give an account of the quasi-isometric rigidity of the mapping class group of a closed orientable surface. Quasi-isometric rigidity was established in [Hamenstädt 2005; Behrstock, Kleiner, Minsky and Mosher 2012]. Here, we give a strengthening of this result which applies to quasi-isometric embeddings (see Theorem 1.1 below).

Many of our arguments have parallels with those of [Behrstock, Kleiner, Minsky and Mosher 2012], though the details are different. Another aim of this paper is to set these arguments in a broader context. The key observation, made in [Bowditch 2013], is that the mapping class group admits a "coarse median" structure. The median in this case is the *centroid* constructed in [Behrstock and Minsky 2011]. Here, in Section 7, we list a set of axioms related to subsurface projection (cf., [Masur and Minsky 2000]) which imply the existence of medians (see Theorem 1.4 below). The point is that the same axioms apply in other situations, notably to Teichmüller space in either the Teichmüller metric or the Weil–Petersson metric. It then follows that these also admit a coarse median structure. This is explained, respectively, in [Bowditch 2016a] and [Bowditch 2015], where various consequences of this observation for the large-scale geometry of these spaces are explored. Again, many of the arguments follow along similar lines, and several general results of this

MSC2010: 20F65.
Keywords: mapping class group, quasi-isometry, rigidity, median.

paper are used in those papers (see, for example, Propositions 1.2 and 1.3 below, as well as the structure of cubes discussed in Sections 10–12).

We begin by outlining the main results of this paper.

Let Σ be a compact orientable surface of genus g with p boundary components. Let $\xi(\Sigma) = 3g + p - 3$ be the *complexity* of Σ. Write $\mathrm{Map}(\Sigma)$ for the mapping class group. When this is viewed as a geometric object, we will use different notation. In particular, we will write $\mathbb{M}(\Sigma)$ for the "marking graph" of Σ as discussed in Section 8. In fact, any proper geodesic space on which $\mathrm{Map}(\Sigma)$ acts isometrically, properly discontinuously and compactly (such the Cayley graph with respect to any finite generating set) would serve for the present discussion. Any two such spaces will be $\mathrm{Map}(\Sigma)$-equivariantly quasi-isometric, by the Schwarz–Milnor Lemma.

It is shown in [Hamenstädt 2005; Behrstock and Minsky 2008] that $\mathbb{M}(\Sigma)$ has coarse rank equal to $\xi(\Sigma)$; that is the maximal dimension ν such that $\mathbb{M}(\Sigma)$ admits a quasi-isometric embedding of $\mathbb{R}^\nu$ (see also [Eskin, Masur and Rafi 2017, Corollary C] and [Bowditch 2013, Theorem 2.6]). Note that it follows that if Σ and Σ' are compact orientable surfaces with $\mathbb{M}(\Sigma)$ quasi-isometric to $\mathbb{M}(\Sigma')$, then $\xi(\Sigma) = \xi(\Sigma')$.

We will show:

Theorem 1.1. *Suppose that Σ and Σ' are compact orientable surfaces with $\xi(\Sigma) = \xi(\Sigma') \geq 4$, and that $\phi : \mathbb{M}(\Sigma) \to \mathbb{M}(\Sigma')$ is a quasi-isometric embedding. Then $\Sigma = \Sigma'$, and ϕ is a bounded distance from the isometry of $\mathbb{M}(\Sigma)$ induced by some element of* $\mathrm{Map}(\Sigma)$.

It immediately follows that ϕ is, in fact, a quasi-isometry. One can also deal, modulo some qualifications, with lower-complexity cases (see the discussion after Theorem 15.2 here). As observed above, if one assumes that ϕ is a quasi-isometry (and that $\Sigma = \Sigma'$), then this statement is given in [Hamenstädt 2005] and [Behrstock, Kleiner, Minsky and Mosher 2012].

We remark that if one assumes quasi-isometric rigidity as given in those papers, then one recovers (indirectly) that the quasi-isometry type of $\mathbb{M}(\Sigma)$ determines the topological type of Σ (modulo a few low-dimensional exceptional cases) since it determines $\mathrm{Map}(\Sigma)$ up to isomorphism (see, for example, [Rafi and Schleimer 2011] for a proof that $\mathrm{Map}(\Sigma)$ determines Σ). Given this, Theorem 1.1 would be equivalent to asserting that any quasi-isometric embedding of $\mathbb{M}(\Sigma)$ into itself is necessarily a quasi-isometry (at least when $\xi(\Sigma) \geq 4$). However, we will give another proof of the rigidity statement in this paper.

As noted above, we base our account around the notion of a coarse median space, as defined in [Bowditch 2013]. This is a geodesic metric space equipped with a ternary operation satisfying certain conditions. Roughly speaking, these say that when dealing with a finite number of points in the space, the ternary operation behaves, up to bounded distance, like the standard median operation on the vertex set of a (finite) CAT(0) cube complex. Such a space comes with a notion of "rank" which is the maximal dimension of such a cube complex needed for the hypothesis.

A related, but different, notion is that of a median metric space, which is also central to our discussion. The definition of a median metric space is quite simple, and is given in Section 2. For further discussion, see [Verheul 1993; Chatterji, Druţu and Haglund 2010; Bowditch 2016c]. This has also been studied from a combinatorial viewpoint; see for example, [Chepoi 2000]. In a median metric space, any triple of points has a unique "median", that is, a point lying between any pair in the triple. This defines a continuous ternary operation, and gives the space the structure of a topological median algebra. (For expositions of the theory of median algebras, see [Isbell 1980; Bandelt and Hedlíková 1983; Roller 1998].) Again, one can associate a "rank" to such a space as the maximal dimension of an embedded cube. (Any CAT(0) cube gives rise example of such a space, after one replaces the euclidean metric on each cube with the l^1 metric. The vertex set is then also such a space.) One can show that a complete connected median metric space of finite rank is canonically bi-Lipschitz equivalent to a CAT(0) metric, [Bowditch 2016c].

The asymptotic cone (see [van den Dries and Wilkie 1984; Gromov 1993]) of a coarse median space is a topological median algebra. If the space has finite rank ν, then the asymptotic cone is bi-Lipschitz equivalent to a median metric space of rank at most ν (see [Behrstock, Druţu and Sapir 2011; Bowditch 2014a] and Theorem 6.9 here). Also, the dimension of any compact subset thereof has dimension at most ν. (This follows from [Bowditch 2013] as we discuss in Section 2.) From the fact that $\mathrm{Map}(\Sigma)$ is a coarse median space one gets a median on its asymptotic cone. This was previously obtained by other means in [Behrstock, Druţu and Sapir 2011]. Much of this is elaborated upon in [Bowditch 2013; 2014a]. Here we obtain more information about the flats in such spaces, which we use for the rigidity result of Theorem 1.1. Similar statements can be found in [Behrstock, Kleiner, Minsky and Mosher 2012], though more specifically for the mapping class group.

We remark that, in [Rafi and Schleimer 2011], the rigidity of the mapping class group is used to deduce the rigidity of the curve graph. Again, it would be interesting to generalise this to quasi-isometric embeddings. As the authors observe, much of their paper works for such embeddings. However there is a key point (aside from their references to [Hamenstädt 2005; Behrstock, Kleiner, Minsky and Mosher 2012]) where an inverse quasi-isometry is needed.

We briefly state a few of the key results proven in this paper, which are used in proving Theorem 1.1, and/or have applications elsewhere.

The first two relate to a median metric space. By a *real cube* in such a space, we will mean a median-convex subset isometric to a finite l^1-product of compact real intervals. (See Section 3 for more precise definitions.) A (closed) subset is *cubulated* if it is a locally finite union of real cubes. We show:

Proposition 1.2. *Suppose that M is a complete median metric space of rank $\nu < \infty$, and that $\Phi \subseteq M$ is a closed subset homeomorphic to $\mathbb{R}^\nu$. Then Φ is cubulated.*

This is proven in Section 4 (see Proposition 4.3). Under additional topological assumptions one can show that Φ is median-convex and isometric to $\mathbb{R}^\nu$ with the l^1 metric (see Proposition 4.6). Using this, one gets a result about products of $\mathbb{R}$-trees:

Proposition 1.3. *Suppose that M is a complete median metric space of rank $\nu < \infty$, with $\nu \geq 2$. Suppose that D is a direct product of ν $\mathbb{R}$-trees, and that none of the factors has a point of valence 2 (i.e., a point which separates the $\mathbb{R}$-tree into exactly 2 components). Suppose that $f : D \to M$ is a continuous injective map, with closed image, $f(D) \subseteq M$. Then f is a median homomorphism, and $f(D)$ is median-convex in M.*

This is shown at the end of Section 4 (Proposition 4.8). This result is used in [Bowditch 2015; 2016a]. A more direct proof in a specific case is given in [Bowditch 2016b] (see Proposition 2.1 thereof). Analogous, but different, statements can be found in [Kleiner and Leeb 1997] and [Kapovich, Kleiner and Leeb 1998].

We make much use of subsurface projections from the marking graph, $\mathbb{M}(\Sigma)$, to curve graphs associated to subsurfaces of Σ. In Section 7, we condense the essential information we need into a set of axioms, (A1)–(A10). This means that much of the argument can be put in a more general setting. In particular, we have the following paraphrasing of a result which will be stated more formally in Section 7; see Theorems 7.1 and 7.2.

Theorem 1.4. *Suppose that to each subsurface, X, of Σ, we have associated geodesic metric spaces, $\mathcal{M}(X)$ and $\mathcal{G}(X)$, together with a collection of projection maps between them satisfying axioms (A1)–(A10). Then each $\mathcal{M}(X)$ has the natural structure of a coarse median space in such a way that each projection map is a quasimorphism (i.e., a median homomorphism up to bounded distance).*

Note that this includes the case where $X = \Sigma$. Here, the spaces $\mathcal{G}(X)$ are (assumed to be) uniformly hyperbolic and the median is the usual centroid in such a space. The various constants involved in the conclusion depend only on those of the hypotheses (A1)–(A10).

In this paper, we are interested mainly in the case where $\mathcal{M}(X)$ and $\mathcal{G}(X)$ are respectively the marking graph, $\mathbb{M}(X)$ and $\mathbb{G}(X)$ and the curve graph of the subsurface X. The same axioms can also be applied to Teichmüller space in either the Teichmüller metric [Bowditch 2016a] or the Weil–Petersson metric [Bowditch 2015].

A simple consequence of Theorem 1.4 is that the asymptotic cone of $\mathcal{M}(\Sigma)$ is a topological median algebra. In fact, it is bi-Lipschitz equivalent to a median metric space, which then allows us to bring the results mentioned above into play.

In outline this paper is structured as follows. Sections 2 to 4 are devoted to a general discussion of median metric spaces. In Section 5 we review properties of asymptotic cones. In Section 6 we discuss general coarse median spaces. In Section 7 we give a set of hypotheses relating to subsurface projection which imply that a geodesic metric space admits a coarse median, and give a precise formulation and proof of Theorem 1.4. This is then applied to the marking graph in Section 8. In Sections 9 to 13, we explore further properties of the marking graph and its asymptotic cone, setting as much as possible in a general context (so that it can be applied elsewhere to Teichmüller space). In Section 14, we explain, in general terms, how the asymptotic cone can be used to control Hausdorff distance. Finally, in Section 15, this applied to the marking complex, to give a proof of Theorem 1.1, together with some discussion of the lower complexity cases.

Notation. Throughout this paper, we will use $\mathbb{G}(X)$ and $\mathbb{M}(X)$ respectively to denote the curve graph and marking graph of a subsurface X of Σ. (We allow $X = \Sigma$, and we need to modify the definitions in the case where X is an annulus, as discussed in Section 8.) We will use the notation $\mathcal{G}(X)$ and $\mathcal{M}(X)$ when the statements apply to the more general spaces satisfying the axioms laid out in Section 7. (This symbol $\mathcal{M}$ will generally denote a coarse median space, as in Section 6.) Note that the curve graph $\mathbb{G}(X)$ plays two slightly different roles: it is one of the family of spaces satisfying these axioms; also its vertex set can be identified with the set of annular subsurfaces of X, and which in this capacity can be viewed as an indexing set. (We remark that in [Bowditch 2013], and some other references, the notation $\Theta(X)$ and $\Lambda(X)$ was used respectively for $\mathcal{G}(X)$ and $\mathcal{M}(X)$.) For the main applications in the present paper, there would be no loss in interpreting $\mathcal{G}(X)$, $\mathcal{M}(X)$ as $\mathbb{G}(X)$, $\mathbb{M}(X)$, respectively.

2. Median metric spaces

We begin with some general discussion of median metric spaces. For elaboration relevant to this paper, see for example, [Verheul 1993; Chatterji, Druţu and Haglund 2010; Bowditch 2016c].

Let (M, ρ) be a metric space. Given $a, b \in M$, let

$$[a, b] = [a, b]_\rho = \{x \in M \mid \rho(a, b) = \rho(a, x) + \rho(x, b)\}.$$

Thus, $[a, b] = [b, a]$ and $[a, a] = \{a\}$.

Definition. We say ρ is a *median metric* if, for all $a, b, c \in M$, $[a, b] \cap [b, c] \cap [c, a]$ consists of exactly one element of M.

We denote this element by $\mu(a, b, c)$ — the *median* of a, b, c. It follows using [Sholander 1954] that (M, μ) is a median algebra (see [Verheul 1993; Chatterji, Druţu and Haglund 2010] and Section 2 of [Bowditch 2016c]). Moreover, $[a, b]$ is exactly the median interval between a and b, i.e., $[a, b] = [a, b]_\mu = \{x \in M \mid \mu(a, b, x) = x\}$. Conversely, note that if (M, μ) is a median algebra, and ρ is a metric satisfying $[a, b]_\mu = [a, b]_\rho$ for all $a, b \in M$, then ρ is a median metric inducing μ. Also, the map $\mu : M^3 \to M$ is continuous (that is, M is a "topological median algebra").

The following definitions only require the median structure on M.

Definition. A subset $B \subseteq M$ is a *subalgebra* if it is closed under μ. It is *convex* if $[a, b] \subseteq B$ for all $a, b \in B$. An *n-cube* is a subset of M median-isomorphic to the direct product of n two-point median algebras: $\{-1, 1\}^n$. (Note that any two-point set admits a unique median structure.) We refer to a 2-cube as a *square*. The *rank* of M is the maximal n such that M contains an n-cube. The rank is deemed to be infinite if there are cubes of all dimensions.

Given $A \subseteq M$ write $\langle A\rangle$ and $\mathrm{hull}(A)$ respectively for the subalgebra generated by A and the convex hull of A, that is, respectively, the smallest subalgebra and smallest convex set in M containing A. Clearly $\langle A\rangle \subseteq \mathrm{hull}(A)$. If A is finite, then so is $\langle A\rangle$. In fact, $|\langle A\rangle| \leq 2^{2^{|A|}}$. Any finite median algebra can be canonically identified as the vertex set of a finite CAT(0) complex; see [Chepoi 2000].

Note that any interval in a connected median metric space is connected (since the map $x \mapsto \mu(a, b, x)$ is a continuous retraction to $[a, b]$). It follows that any connected component of a median metric space is convex.

Definition. We say that a median metric space is *proper* if it is connected, complete and has finite rank.

Henceforth we will assume that M is a proper median metric space, though as we will comment, many of the constructions only require it to be a median metric space, or indeed just a median algebra.

It was shown in [Bowditch 2014a, Corollary 1.3] that if M is proper, then every interval $[a, b]$ in M is compact. (One can go on to deduce that the convex hull of any compact set is compact.)

We say that a topological median algebra is *locally convex* if every point has a base of convex neighbourhoods.

Lemma 2.1. *Any median metric space M of finite rank is locally convex.*

Proof. This follows since M is "weakly locally convex" in the sense of [Bowditch 2013, Section 7]. (Note that if $a, b \in M$, then the diameter of $[a, b]$ is equal to $\rho(a, b)$.) Since it has finite rank, Lemma 7.1 of [loc. cit.], tells us that it is locally convex. □

(In the case of interest here, namely the asymptotic cone of a finite rank coarse median space, the conclusion also follows from Lemma 9.2 of [loc. cit.].)

It was also shown there (Theorem 2.2) that any locally compact subset of M has topological dimension at most $\mathrm{rank}(M)$. (For more discussion of dimension, see Section 4 of the present paper.)

Theorem 2.2 [Bowditch 2016c, Theorem 1.1]. *If (M, ρ) is a proper median metric space, then there is a canonically associated bi-Lipschitz equivalent metric, σ_ρ, on M for which (M, σ_ρ) is CAT(0).*

In fact, we can arrange that $\rho/\sqrt{\mathrm{rank}(M)} \le \sigma_\rho \le \rho$.

Note that it immediately follows that M is contractible.

A simple example is $\mathbb{R}^n$ with the l^1 metric. In this case, σ_ρ recovers the euclidean metric on $\mathbb{R}^n$. Any convex subset of $\mathbb{R}^n$ has the form $P = \prod_{i=1}^n I_i$ where $I \subseteq \mathbb{R}$ is a real interval (possibly unbounded). If each I_i is either a singleton or all of $\mathbb{R}$, we refer to P as a *coordinate plane*. If each $I_i = [a_i, b_i]$ with $a_i < b_i$, we refer to P as an l^1 *cube*. We refer to $P = \prod_{i=1}^n (a_i, b_i)$ as the *relative interior* of P, and we refer to the elements of $Q = \prod_{i=1}^n \{a_i, b_i\}$ as the *corners* of P. Note that these are determined by the intrinsic geometry of P. Also $P = \mathrm{hull}(Q)$. In fact, $P = [a, b]$ where a, b are any pair of opposite corners of P.

Another class of examples arise from CAT(0) complexes. Suppose that Υ is (the topological realisation of) a finite CAT(0) complex. Suppose that each cell is given the structure of an l^1-cube. This induces a path metric, ρ, on Υ, so that (Υ, ρ) is a median metric space. In this case, (Υ, σ_ρ) is a euclidean CAT(0) cube complex, where we can allow the cells to be rectilinear parallelepipeds.

Definition. We refer to a space of the form (Υ, ρ) as an l^1-*cube complex*.

There is a sense in which any proper metric median space can be approximated by subspaces of this form. The following was shown in [Bowditch 2016c, Lemmas 7.5 and 7.6].

Lemma 2.3. *Let (M, ρ) be a complete connected median metric space. Suppose that $\Pi \subseteq M$ is a finite subalgebra. Then there is a closed subset $\Upsilon \subseteq M$ which has the structure of a finite l^1-cube complex in the induced metric ρ, and such that $\Pi \subseteq \Upsilon$ is exactly the set of vertices of this complex.*

The statement is taken to imply that the metric ρ restricted to Υ is already a path metric on Υ. In general, Υ will not be unique. (One can make a canonical choice by taking cells to be totally geodesic in the metric σ_ρ on M, but we will not need this here.) Note that we do not assume here that the cells of Υ are convex in M. (If that were the case, we refer to Υ as a "straight" cube complex, as we will define more formally in Section 3.)

We continue with some more general observations. For the moment, M can be any median metric space.

Given $a, b \in M$, we define $\phi = \phi_{a,b} : M \to [a, b]$ by $\phi(x) = \mu(a, b, x)$. This is a 1-Lipschitz median epimorphism.

Definition. We say that two pairs (a, b), (c, d) in M^2 are *parallel* if $[b, c] = [a, d]$.

It is equivalent to saying that both $b, c \in [a, d]$ and $a, d \in [b, c]$. When a, b, c, d are all distinct, it is also equivalent to saying that a, b, d, c is a square. Note that parallelism is an equivalence relation on M^2. If a, b and c, d are parallel, then $\phi_{a,b}|[c, d]$ is an isometry (hence a median isomorphism) from $[c, d]$ to $[a, b]$. Its inverse is $\phi_{c,d}|[a, b]$.

The following is a standard notion for median algebras.

Definition. If $C \subseteq M$ is closed and convex, we say that $\phi : M \to C$ is a *gate map* of M to C if $\phi(x) \in [x, c]$ for all $x \in M$ and $c \in C$.

A more detailed discussion of gate maps can be found in [Bowditch 2016a, Section 2.4].

One verifies that ϕ is a 1-Lipschitz retraction of M to C, and a median homomorphism. If ϕ exists then it is unique. Note that the map $\phi_{a,b}$ of the previous paragraph is a gate map to $[a, b]$. In fact, if M is proper, then gate maps to closed convex sets always exist. This can be seen using the fact that intervals are compact, though we will not need this here.

Definition. A *wall* in M is a partition of M into two nonempty convex subsets.

This is equivalent to a median epimorphism $\phi : M \to \{-1, 1\}$, where the partition is given by $\{\phi^{-1}(-1), \phi^{-1}(1)\}$. We can speak about an *oriented* or *unoriented* wall according to whether we consider the partition as an ordered or an unordered pair. Any two disjoint convex subsets, C, D, of M are separated by some wall, that is, $C \subseteq \phi^{-1}(-1)$ and $D \subseteq \phi^{-1}(1)$. We say two walls, ϕ, ψ, *cross* if the map $(\phi, \psi) : M \to \{-1, 1\}^2$ is surjective. The rank of M can be equivalently defined as the maximal cardinality of a set of pairwise crossing walls. We say that M is *n-colourable* if we can colour the walls of M with n colours such that no two walls of the same colour cross. This implies that $\text{rank}(M) \le n$. (See [Bowditch 2013, Section 12] for more discussion of colourability.)

These notions only require the median structure on M. If Π is a finite median algebra, then we can identify the set of (unoriented) walls with the set of hyperplanes in the associated finite CAT(0) complex. In this case, two walls cross if and only if the corresponding hyperplanes intersect.

If $a, b \in M$, then $[a, b]$ admits a partial order defined by $x \le y$ if $x \in [a, y]$ (or equivalently $y \in [x, b]$). If $[a, b]$ has rank 1, this is a total order. If M is connected and metrisable, then $[a, b]$ is isometric to a compact real interval. In particular, any connected median metric space of rank 1 is an $\mathbb{R}$-tree. (In this case, the metric σ_ρ, described above, agrees with ρ.)

We also note the following construction of quotient median algebras. Suppose that M is a median algebra, and that $\sim$ is an equivalence relation on M such that whenever $a, b, c, d \in M$ with $c \sim d$, then $\mu(a, b, c) \sim \mu(a, b, d)$. Let $P = M/\sim$. Given $x, y, z \in P$, set $\mu_P(x, y, z)$ to be the equivalence class of $\mu(a, b, c)$, where a, b, c are representatives of x, y, z respectively. This is well defined, the quotient (P, μ_P) is a median algebra, and the quotient map is an epimorphism. Indeed any epimorphism of median algebras arises in this way.

We finish this section with the following proposition which will be applied to asymptotic cones of finite-rank coarse median spaces (see Lemma 6.6).

Proposition 2.4. *Let (M, μ) be a median algebra with* $\mathrm{rank}(M) \leq \nu$, *and let ρ be a geodesic metric on M. Suppose that there is some $k \geq 1$ such that for all $a, b, c, d \in M$, we have $\rho(\mu(a, b, c), \mu(a, b, d)) \leq k\rho(c, d)$. Then there is a median metric λ on M, bi-Lipschitz equivalent to ρ, and which induces the median μ. Moreover, the bi-Lipschitz constants depend only on ν and k.*

The second hypothesis asserts that the projection to intervals is uniformly Lipschitz. (It is precisely axiom (L2) of [Bowditch 2014a, Section 1].) It implies that the median operation is Lipschitz hence continuous (so M is a topological median algebra). In fact, we can weaken the geodesic condition to assert that M is Lipschitz path-connected, in the sense of axiom (L3) there. The proof is the same, but we won't need the more general statement here.

In the same work, it was shown that if (M, μ) is also finitely colourable, then it embeds in a finite product of trees, so the induced metric is median. The proof below amounts to observing that, under the weaker hypothesis of finite rank, the same construction gives a median metric directly.

Proof of Proposition 2.4. We write $\langle a, b{:}c\rangle_\lambda = \frac{1}{2}(\lambda(a, c) + \lambda(b, c) - \lambda(a, b))$ (i.e., the "Gromov product"). Then $[a, b]_\lambda = \{x \in M \mid \langle a, b{:}x\rangle_\lambda = 0\}$. Thus, λ is a median metric inducing μ if and only if $[a, b]_\lambda = [a, b]$ for all $a, b \in M$. In this case, given any $a, b, c \in M$ we have $\rho(c, d) = \langle a, b{:}d\rangle_\lambda$, where $d = \mu(a, b, c)$ (see, for example, [Bowditch 2014a, Section 2]).

Now, if $\Pi \subseteq M$ is a finite subalgebra, we put a metric, λ_Π, on Π as in [Bowditch 2014a, Section 5]. (To each wall, W, of Π we associate a "width", $\lambda(W)$, and set $\lambda_\Pi(a, b) = \sum_W \lambda(W)$, as W ranges over the set, $\mathcal{W}(a, b)$, of walls separating a from b.) It is easily seen that $[a, b]_{\lambda_\Pi} = [a, b] \cap \Pi$, the latter being the intrinsic median interval in Π. (This holds since $\mathcal{W}(a, b) \subseteq \mathcal{W}(a, c) \cup \mathcal{W}(b, c)$, with equality if and only if $c \in [a, b] \cap \Pi$.) Therefore λ_Π is a median metric on Π.

Moreover, λ_Π is uniformly bi-Lipschitz equivalent to ρ restricted to Π. This follows as in [Bowditch 2014a, Section 5]. Note that if $x, y \in M$, then $[x, y] \cap \Pi$ has rank at most ν. It follows by Dilworth's lemma that $[x, y] \cap \Pi$ is ν-colourable (Lemma 2.3 of the same work) and so embeddable as a subalgebra of the cube

$[0, 1]^\nu$ (Proposition 1.4 there). Lemmas 5.2 and 5.4 there then respectively give us lower and upper bounds on $\lambda(x, y)$ in terms of $\rho(x, y)$. (Note that, in the notation of that work, $T \leq \rho(x, y)$, if we assume that M is a geodesic space.)

As in Section 6 of [Bowditch 2014a], we note that the set of finite subalgebras of M, ordered by inclusion, is cofinal in the set of all finite subsets of M. Therefore, by Tychonoff's theorem, we find a cofinal set of finite subalgebras, Π, so that $\lambda_\Pi(a, b) \to \lambda(a, b)$ for all $a, b \in M$, where λ is a metric on M, bi-Lipschitz equivalent to ρ.

To see that λ is a median metric inducing ρ, we need to check that $[a, b]_\lambda = [a, b]$. To this end, suppose $a, b, c \in M$ and let $d = \mu(a, b, c)$. Note that $a, b, c, d \in \Pi$ for a cofinal subset of those Π in our cofinal set of subalgebras. If $c \in [a, b]$, then $\langle a, b{:}c\rangle_{\lambda_\Pi} = 0$ for all such Π, so $\langle a, b{:}c\rangle_\lambda = 0$, so $c \in [a, b]_\lambda$. Conversely, if $c \in [a, b]_\lambda$, then $\lambda_\Pi(c, d) = \langle a, b{:}c\rangle_{\lambda_\Pi} \to 0$, so $\lambda(c, d) = 0$, so $c = d$, so $c \in [a, b]$. □

3. Blocks

In this section, we describe top-dimensional cubes in median metric spaces.

Let M be a proper median metric space. Throughout this section, we will use ν to denote $\mathrm{rank}(M)$.

Definition. An *n-block* in M is a convex subset isometric to an l^1- product of n nontrivial compact real intervals.

This is equivalent to saying that it is convex and median-isomorphic to $[-1, 1]^n$. Clearly, $n \leq \nu$.

We write $P \equiv \prod_{i=1}^n I_i$, where each I_i is a compact real interval, and can be identified with a 1-face of P.

Let $Q(P)$ be the set of corners of P, that is, $Q(P) = \prod_i \{a_i, b_i\}$ where $I_i = [a_i, b_i]$. It is clear that $Q(P)$ is intrinsically an n-cube in P, hence an n-cube in M. We see $P = \mathrm{hull}(Q(P))$. In fact, $P = [a, b]$, where a, b are any pair of opposite corners of Q.

Lemma 3.1. *Let M be a proper median metric space of rank ν. The following are equivalent for a subset $P \subseteq M$:*

(1) *P is ν-block.*

(2) *P is the convex hull of a ν-cube in M.*

(3) *P is isometric to a ν-dimensional l^1-cube.*

Proof. The fact that (2) implies (1) was proven in [Bowditch 2016c, Proposition 5.6]. Suppose (3) holds. Let a, b be opposite corners of P (defined intrinsically). Directly from the definition of intervals in M, we can see that $P \subseteq [a, b]$, and so $P \subseteq \mathrm{hull}(Q)$, where Q is the set of corners of P. By the observation preceding the lemma, we know that $\mathrm{hull}(Q)$ is a ν-block, and it now follows easily that we must have $P = \mathrm{hull}(Q)$. □

In (3) here, we are assuming that P is isometric to an l^1-cube in the induced metric. We suspect that it would be sufficient to assume that this were the case for the induced path-metric. We will show this to be the case under some regularity assumptions (see Lemma 3.4 below).

Lemma 3.2. *Let M be a proper median metric space of rank ν. Suppose that P, $P' \subseteq M$ are ν-blocks, and that $P \cap P'$ is a common codimension-1 face. Then $P \cup P'$ is also a ν-block.*

Proof. Let $R_0 = Q(P \cap P') = Q(P) \cap Q(P')$. Let $R = Q(P) \setminus R_0$ and $R' = Q(P') \setminus R_0$. Thus R_0, R, R' are parallel $(\nu-1)$-cubes. In particular, $R \cup R'$ is a ν-cube. Let $P'' = \mathrm{hull}(R \cup R')$. By Lemma 3.1, this is a ν-block. We claim that $R_0 \subseteq P''$. For if $r_0 \in R_0$, let $r \in R$ and $r' \in R'$ be adjacent vertices of $Q(P)$ and $Q(P')$ respectively. Thus, $[r_0, r]$ and $[r_0, r']$ are 1-faces of $Q(P)$ and $Q(P')$. In particular, $[r_0, r] \cap [r_0, r'] = \{r_0\}$ and so $r_0 \in [r, r'] \subseteq P''$ as claimed. It now follows that $P \cup P' = P''$. □

More generally, if P, P' are any two blocks, then so is $P \cap P'$ provided it is nonempty. In fact, $P \cap P' = \mathrm{hull}(Q)$, where Q is the projection (image of the gate map) of $Q(P')$ to P. In particular, $Q(P \cap P') \subseteq \langle Q(P) \cup Q(P') \rangle$.

We have the following procedure for subdividing blocks. Suppose $P \equiv \prod_{i=1}^n I_i$. If $F_i \subseteq I_i$ are finite subsets containing the endpoints, then $F = \prod_{i=1}^n I_i$ is a finite subalgebra of P. In fact, any finite subalgebra of P containing Q has this form. We can represent P as an l^1-cube complex whose vertex set is exactly F. We refer to this as a *subdivision* of P.

Lemma 3.3. *Suppose that $\mathcal{P}$ is a finite set of blocks in M. Then we can subdivide these blocks to find another set of blocks, $\mathcal{P}'$, with $\bigcup \mathcal{P} = \bigcup \mathcal{P}'$ such that any two blocks of $\mathcal{P}'$ meet, if at all, in a common face.*

Proof. Let $A = \bigcup_{P \in \mathcal{P}} Q(P)$ and let $\Pi = \langle A \rangle$. If $P \in \mathcal{P}$, then $P \cap \Pi$ is a subalgebra of P containing $Q(P)$ and so determines a subdivision of P. We subdivide each element of $\mathcal{P}$ in this way to give us our new collection $\mathcal{P}'$. Now if P, $P' \in \mathcal{P}'$, then $Q(P \cap P') \subseteq \langle Q(P) \cup Q(P') \rangle \subseteq \Pi$. But by construction, $P \cap \Pi \subseteq Q(P)$ and $P' \cap \Pi \subseteq Q(P')$, so $Q(P \cap P') \subseteq P \cap P' \cap \Pi \subseteq Q(P) \cap Q(P')$. It now follows that $P \cap P'$ is a common face of P and P' as claimed. □

In other words, we can realise $\bigcup \mathcal{P}$ as an l^1-cube complex in M all of whose cells are blocks.

Definition. A *straight cube complex* in M is an embedding of a locally finite cube complex in M such that each cell is a block (necessarily of the corresponding dimension).

The following is equivalent to the informal definition of "cubulated" given in Section 1.

Definition. A *cubulated set* is a subset of M which is a locally finite union of blocks.

A cubulated set, Φ, is clearly closed, and by the above, we see that any point $x \in \Phi$ has a neighbourhood in Φ which is a straight cube complex contained in Φ. In fact, we can assume that x is a vertex of this cube complex. Note also that a finite union or a finite intersection of cubulated sets is also cubulated.

In fact, if $\Phi_1, \ldots, \Phi_n$ is a finite set of cubulated sets, with $x \in \bigcap_i \Phi_i$, then we can find a straight cube complex, $\Upsilon \subseteq \bigcup_i \Phi_i$ as above, with each $\Upsilon \cap \Phi_i$ a subcomplex of Υ. (This is a consequence of the construction of Lemma 3.3.)

Lemma 3.4. *Suppose that $\Phi \subseteq M$ is cubulated. Suppose that $P \subseteq \Phi$ is isometric to a ν-dimensional l^1-cube in the path-metric induced from ρ. Then P is a ν-block in M.*

Proof. By Lemma 3.3 we can find a straight cube complex $\Upsilon \subseteq \Phi$, with $P \subseteq \Upsilon$. We can assume that the intrinsic corners of P are all vertices of Υ. It now follows that P is a union of ν-blocks of M, which are ν-cells of Υ. These determine a subdivision of P in the induced path metric on P. Applying Lemma 3.2 inductively, we see that P is a block in M. □

Definition. Suppose that $\Phi \subseteq M$ is cubulated. We say that a point $x \in \Phi$ is *regular* if it has a neighbourhood in Φ which is a ν-block in M. Otherwise, we say that x is *singular*. We write Φ_S for the set of singular points of Φ.

Note that Φ_S is a cubulated set of dimension at most $\nu - 1$.

Suppose now that Φ is cubulated and homeomorphic to $\mathbb{R}^\nu$. If $K \subseteq \Phi$ is compact, then K lies inside a straight cube complex, Υ, in Φ. Moreover, we can assume that any $(\nu-1)$-cell of Υ meeting K lies in exactly two ν-cells of Υ. By Lemma 3.2, the union of these to cells is also a ν-block in M. From this, we deduce:

Lemma 3.5. *Suppose that $\Phi \subseteq M$ is cubulated and homeomorphic to $\mathbb{R}^\nu$. Then Φ_S is a cubulated set of dimension at most $\nu - 2$.*

Note that, if P is any block in Φ, then the relative interior of P in Φ is exactly the intrinsic relative interior of P, as defined earlier.

Definition. A *leaf segment* of Φ is a closed subset, L, of Φ homeomorphic to a real interval such that if $x \in L$, then there is a block $P \subseteq \Phi$ containing x in its relative interior, with $L \cap P$ lying in a coordinate line of P. If the real interval is the whole real line, we refer to L as a *leaf*.

Clearly this implies that $L \cap \Phi_S = \emptyset$. We note:

Lemma 3.6. *Every leaf segment of Φ is convex in M.*

Proof. Let $L \subseteq \Phi$ be a leaf segment, and suppose $I \subseteq L$ is a compact subinterval. Since $I \cap \Phi_S = \varnothing$, we can find a subset $P \subseteq \Phi$ which is a block in the intrinsic path metric on P, and with $I \subseteq P$ an intrinsic coordinate line with respect to that structure. But by Lemma 3.4, P is a block in M, and so I is convex. It now follows that L is convex. □

Definition. A *flat* in M is a closed convex subset isometric to $\mathbb{R}^\nu$ with the l^1 metric.

(Note that we always take a flat to be of maximal dimension; that is, $\nu = \mathrm{rank}(M)$.)

In fact (as with blocks), we see that any closed subset of M which is isometric to $\mathbb{R}^\nu$ in the induced metric is flat. (Indeed, we suspect this remains true if we substituted "induced path-metric" for "induced metric" in the above.) Also, any closed convex subset of M median isomorphic to $\mathbb{R}^\nu$, with the standard product structure, is a flat. In particular, the notion depends only on the topology and median structure.

Clearly a flat is a cubulated set with empty singular set. Conversely, we have:

Lemma 3.7. *Suppose that $\Phi \subseteq M$ is a cubulated set homeomorphic to $\mathbb{R}^\nu$, and with $\Phi_S = \varnothing$. Then Φ is a flat.*

Proof. First note that, in the intrinsic path metric, Φ is locally isometric to $\mathbb{R}^\nu$ in the l^1 metric. Since it is complete, it must be globally isometric. By Lemma 3.4 any subset of Φ that is intrinsically a block is indeed a block in M, and so, in particular, convex. Since any two points of Φ are contained in such a subset, it follows that Φ is convex. The induced path metric is therefore the same as the induced metric. □

Here is a criterion for recognising that a cubulated set is indeed nonsingular:

Lemma 3.8. *Suppose that $\Phi \subseteq M$ is cubulated, and that there is a homeomorphism $f : \mathbb{R}^\nu \to \Phi$ such that if $H \subseteq \mathbb{R}^\nu$ is any codimension-1 coordinate plane in $\mathbb{R}^\nu$ then $f(H)$ is cubulated. Then Φ is a flat, and f is a median isomorphism.*

Proof. Suppose that $L \subseteq \mathbb{R}^\nu$ is a coordinate line, and that $x \in L$ with $f(x) \notin \Phi_S$. Let $H_1, H_2, \ldots, H_n$ be the codimension-1 coordinate planes through x, with $L = \bigcap_{i=2}^n H_i$, and with H_1 orthogonal to L. As noted after Lemma 3.3, we can find a neighbourhood, Υ, of $f(x)$ in Φ, which is a straight cube complex, with $f(x)$ a vertex, and each $f(H_i) \cap \Upsilon$ a subcomplex of Υ. In particular, $f(L) = \bigcap_{i=2}^n f(H_i)$ is a 1-dimensional subcomplex, and so meets $f(x)$ in a pair of 1-cells of Υ. Let Δ be the link of $f(x)$ in Υ. Since $f(x) \notin \Phi_S$, this is a cross polytope. Note that $f(L)$ determines two vertices, p, q, of Δ. Now $f(H_1)$ separates the two rays of $f(L)$ with basepoint $f(x)$ in Φ. It therefore determines a subcomplex of Δ separating p from q in Δ. It follows that p and q must be opposite vertices of Δ. We see that the union of the two 1-cells of $f(L)$ meeting x is convex.

In summary, we have shown that, away from Φ_S, the images of coordinate lines are locally convex, that is, leaf segments of Φ. By a simple compactness argument, it now follows that if $I \subseteq \mathbb{R}^\nu$ is a compact interval lying in a coordinate line with $f(I) \cap \Phi_S = \varnothing$, then $f(I)$ is a leaf segment of Φ. We can now deduce that if $P \subseteq \Phi$ is any ν-block in $\Phi \setminus \Phi_S$, then $f^{-1}|P$ is a median isomorphism to a block $f^{-1}(P)$ in $\mathbb{R}^\nu$. In fact, it is enough that P should not meet Φ_S in its relative interior.

Suppose now that $y \in \Phi$. Let $\Upsilon \subseteq \Phi$ be a straight cube complex that is a neighbourhood of y, and with y as a vertex. To simplify notation, suppose that $f^{-1}(y)$ is the origin in $\mathbb{R}^\nu$. Let P be a cube of Υ with y a corner of P. Then $f^{-1}P$ has the form $\prod_{i=1}^{\nu}[0, \pm t_i]$ for some $t_i > 0$. Since there are only finitely many such P, after shrinking them, we can assume that the preimages all have the form $\prod_{i=1}^{\nu}[0, \pm t]$ for some $t > 0$. We see that there are exactly 2^ν such cubes, which fit together into a bigger cube of the form $f([-t, t]^\nu)$. In particular, the link of y in Υ is a cross polytope, and y is regular.

We have shown that $\Phi_S = \varnothing$, and so by Lemma 3.7, Φ is a flat. □

For reference elsewhere (see Proposition 4.8 below) we note that there is a variation on Lemma 3.8, where $\mathbb{R}^\nu$ is replaced by a real cube, $[-1, 1]^\nu$, and $\Phi = f([-1, 1]^\nu)$. In fact, it is enough to assume that the relative interior, $f((-1, 1)^n) \subseteq \Phi$, is cubulated (in the sense that any compact subset of $f((-1, 1)^\nu)$ lies inside another compact subset of Φ which is cubulated). Also, we only need to consider coordinate planes restricted to the interior of Φ. The argument is essentially the same, this time applied to the relative interior of Φ and then taking the closure.

4. Cubulating planes

In this section, we discuss the regularity of "top-dimensional manifolds" in M. These play an important role in [Kleiner and Leeb 1997; Kapovich, Kleiner and Leeb 1998; Behrstock, Kleiner, Minsky and Mosher 2012]. Our argument is analogous to those to be found there, though set in a somewhat different context. Here, we interpret this in terms of cubulations. We will only use dimension of (locally) compact sets, so all the standard definitions are equivalent. For definiteness, we can interpret the dimension of a topological space to be its covering dimension. (Note that this differs from the notion of "topological rank" used in [Kleiner and Leeb 1997].)

Suppose that (M, ρ) is a complete median metric space. We first note:

Lemma 4.1. *Any locally compact subset of M has topological dimension at most* $\operatorname{rank}(M)$.

Proof. First note that by Lemma 2.1, M is locally convex. The statement then follows by Theorem 2.2 and Lemma 7.6 of [Bowditch 2013]. □

From this we see that if M is homeomorphic to $\mathbb{R}^\nu$, then $\nu = \operatorname{rank}(M)$. (The fact that $\nu \geq \operatorname{rank}(M)$ is an immediate consequence of Lemma 4.1. For the other direction, note that by Lemma 2.3, any n-cube in M is the vertex set of an embedded l^1-cube in M, and so $n \le \nu$, and it follows that $\operatorname{rank}(M) \le \nu$.)

In fact, we can say a lot more about the regularity of such a space:

Lemma 4.2. *If M is a complete median metric space homeomorphic to $\mathbb{R}^\nu$, then M is cubulated.*

In particular, we see that M is locally isometric to $\mathbb{R}^\nu$ with the l^1 metric away from a cubulated singular set of dimension at most $\nu - 2$. (Note that we are not claiming that the cubulation is combinatorial in the sense of PL manifolds. Certainly the link of any cell in the cubulation will be a homology sphere. It is not clear whether it need be a topological sphere in this situation.)

Proof of Lemma 4.2. Let $B_1 \subseteq B_0$ be topological ν-balls in M. We suppose that $N(B_1; 2u) \subseteq B_0$, where $N(\cdot\,; r)$ denotes the metric r-neighbourhood with respect to the metric ρ. Let $0 < s < t < u$ be sufficiently small depending on u, as described below. We take a topological triangulation of ∂B_0, all of whose simplices have diameter at most s. Let $A \subseteq \partial B_0 \subseteq M$ be the set of vertices of this triangulation, and let $\Pi = \langle A \rangle \subseteq M$. By Lemma 2.3, Π is the vertex set of an l^1-cube complex Υ embedded in M. We extend the inclusion of A into Υ to a continuous map $f : \partial B_0 \to \Upsilon$. Provided s is small enough in relation to t, we can arrange that the ρ-diameter of the image of each simplex is at most t. (For example, take the corresponding euclidean metric, σ_Υ, on Υ. Then $(\Upsilon, \sigma_\Upsilon)$ is CAT(0), and we can map in simplices, inductively on the 1-skeleta by taking geodesic rulings. In this way the σ_Υ-diameter of the image of any simplex is at most s. Now $\rho \le \sigma_\Upsilon \sqrt{\nu}$, so this works provided $s\sqrt{\nu} \le t$.) Now $\rho(x, f(x)) \le s + t$ for all ∂B_0. Again, provided t is small enough in relation to u, we can find a homotopy, $F : \partial B_0 \times [0, 1] \to M$, between f and the inclusion of B_0 into M whose trajectories all have length at most u. In particular, the image of the homotopy lies in $N(\partial B_0; u)$ and is therefore disjoint from B_1. For this, it is convenient to take the CAT(0) metric, σ, on M, as given by Theorem 2.2. We can then use linear isotopy in this metric, that is, the trajectory from x to $f(x)$ is the σ-geodesic segment. Again we note that $\rho \le \sigma\sqrt{\nu}$, so this works provided $(s + t)\sqrt{\nu} \le u$.

Now Υ is a CAT(0) complex in the euclidean metric, and so in particular is contractible. We can therefore extend $f : \partial B_0 \to \Upsilon$ arbitrarily to a continuous map $f : B_0 \to \Upsilon$. We combine this with the homotopy constructed above to give a continuous map $g : B_0 \to M$ which restricts to inclusion on ∂B_0. More formally, if $x \in B_0 \setminus \{0\}$, write $x = \lambda\hat{x}$, where $\lambda \in (0, 1]$ and $\hat{x} \in \partial B_0$ (via any homeomorphism of B_0 with the unit ball in $\mathbb{R}^\nu$). If $\lambda \le \frac{1}{2}$, then set $g(x) = f(2\lambda\hat{x})$. If $\lambda \ge \frac{1}{2}$, then set $g(x) = F(\hat{x}, 2\lambda - 1)$. We set $g(0) = 0$. Note that $g(B_0) = f(B_0) \cup \operatorname{image}(F)$, and we have noted that $B_1 \cap \operatorname{image}(F) = \emptyset$, and so $B_1 \cap g(B_0) \subseteq f(B_0)$. But now, $B_0 \subseteq g(B_0)$ (since $g|\partial B_0$ is just inclusion). It therefore follows that $B_1 \subseteq f(B_0) \subseteq \Upsilon$.

We do not know a priori that Υ is a straight complex. However, every ν-cell of Υ must be a ν-block. Moreover, B_1 must lie in the union of these ν-cells. (For if $x \in B_1$, then any cell of Υ must lie in a ν-cell, otherwise some neighbourhood of x in B_1 would have dimension at most $\nu - 1$.) Since B_1 was an arbitrary ν-ball in M, we see that every compact subset of M lies in a finite union of ν-blocks of M. It follows that M is cubulated. □

We can give a more general version of this for subsets of a proper median metric space as follows (given as Proposition 1.2 in Section 1).

Proposition 4.3. *Suppose that M is a complete median metric space of rank at most ν, and that $\Phi \subseteq M$ is a closed subset homeomorphic to $\mathbb{R}^\nu$. Then Φ is cubulated.*

Clearly, in this case, the rank will be exactly ν. As before, we see that Φ is locally isometric to $\mathbb{R}^\nu$ in the l^1 metric away from a codimension-2 singular set (see Lemma 3.5).

Note that there is no loss in assuming that M is connected (hence "proper" in the terminology of Section 3) since we can simply restrict to the component containing Φ. We have already observed in Section 2 that this is convex, hence intrinsically a complete median metric space.

For the proof, will need the following two topological lemmas:

Lemma 4.4. *Suppose that X is a Hausdorff topological space and that $B, P \subseteq X$ are embedded topological n-balls, with intrinsic boundary spheres $S(B)$ and $S(P)$ respectively. Suppose that $P \setminus S(P)$ is open in X, that $P \cap S(B) = \varnothing$ and that $B \cap P \setminus S(P) \neq \varnothing$. Then $P \subseteq B$.*

Proof. Write $I(B) = B \setminus S(B)$ and $I(P) = P \setminus S(P)$ for the relative interiors. These are both homeomorphic to $\mathbb{R}^n$. Let $U = I(P) \cap B = I(P) \cap I(B)$. By assumption, $U \neq \varnothing$. Now $I(P)$ is open in X, so U is open in $I(B)$. Thus, U is homeomorphic to an open subset of $\mathbb{R}^n$, hence, by invariance of domain, it is also open in $I(P)$. But $U = I(P) \cap B$, so U is also closed in $I(P)$, and so, by connectedness, $U = I(P)$. In other words, $I(P) \subseteq I(B)$, and it follows that $P \subseteq B$ as claimed. □

For the second topological lemma, we need the following definition.

Definition. The *(locally) compact dimension* of a Hausdorff topological space is the maximal topological dimension of any (locally) compact subset.

Clearly the compact dimension is at most the locally compact dimension, which in turn is at most the "separation dimension" as defined in [Bowditch 2013, Section 7].

Lemma 4.5. *Suppose that M is a Hausdorff topological space of compact dimension at most ν. Suppose that B is a topological ν-ball with boundary ∂B. Suppose that $f_0, f_1 : B \to M$ are continuous and homotopic relative to ∂B, and that f_0 is injective. Then $f_0(B) \subseteq f_1(B)$.*

The proof is based on an argument in Section 6.1 of [Kleiner and Leeb 1997]. A related, but slightly different statement can be found in [Behrstock, Kleiner, Minsky and Mosher 2012, Section 6]. In what follows, H_r will denote Čech homology with coefficients in a field (say $\mathbb{Z}_2$ to be specific). We will only deal with compact spaces, so that the usual homology axioms, in particular, homotopy, excision and exactness, hold. We need compact spaces and field coefficients for exactness; see [Eilenberg and Steenrod 1952, Chapter IX]. (Note that in [Kleiner and Leeb 1997], it is implicit from context that singular homology is being used. As a consequence they use open sets instead of compact sets.) Note that, if K is compact and of dimension at most ν, then $H_n(K, A)$ is trivial for any compact $A \subseteq K$ and any $n > \nu$.

Proof. Let $C = f_0(B)$, $D = f_1(B)$, $S = f_0(\partial B) = f_1(\partial B)$ and let $E \subseteq M$ be the image of a homotopy from f_0 to f_1. Thus, $S \subseteq C \cap D \subseteq C \cup D \subseteq E$ are all compact. Suppose, for contradiction, that $p \in C \setminus D$. Let $N \subseteq C$ be an open neighbourhood of p in C, whose closure is homeomorphic to a closed ν-ball disjoint from D. Now $H_\nu(C, C \setminus N) \cong H_{\nu-1}(S) \cong \mathbb{Z}_2$, but the image of $H_\nu(C, C \setminus N)$ in $H_\nu(E, C \cup D \setminus N)$ is trivial. (Note that this corresponds to the image of $H_{\nu-1}(\partial B)$ under that map induced by $f_1 \simeq f_0$.) Now the natural map $H_\nu(C, C \setminus N) \to H_\nu(C \cup D, C \cup D \setminus N)$ is an isomorphism, by excision. Also, since $H_{\nu+1}(E, C \cup D)$ is trivial, the exact sequence of triples tells us that the natural map, $H_\nu(C \cup D, C \cup D \setminus N) \to H_\nu(E, C \cup D \setminus N)$ is injective. Composing, we get that the natural map $H_\nu(C, C \setminus N) \to H_\nu(E, C \cup D \setminus N)$ is injective, giving a contradiction. □

We can now give the proof of Proposition 4.3. We have already observed that we can assume M to be connected. We recall that M is contractible (see Theorem 2.2), and has locally compact dimension at most ν (Lemma 4.1).

Proof of Proposition 4.3. This is an extension of the argument for Lemma 4.2. This time, we take three closed topological balls, $B_2 \subseteq B_1 \subseteq B_0 \subseteq \Phi \subseteq M$. We assume that B_2 is contained in the relative interior of B_1, and that $N(B_1; 2u) \subseteq B_0$ (in the metric ρ on M). We start as before, triangulating ∂B_0, to give us a complex $\Upsilon \subseteq M$, a continuous map $f : B_0 \to \Upsilon$, and a homotopy in M from $f|\partial B_0$ to the inclusion of ∂B_0. We can arrange that the homotopy does not meet B_1. We combine f with this homotopy to give a continuous map, $g : B_0 \to M$, which restricts to the identity on ∂B_0.

Since M is contractible, g is homotopic to the inclusion of B_0 in M, relative to ∂B_0. Therefore, Lemma 4.5 tells us that $B_0 \subseteq g(B_0)$. Moreover, as observed above, the homotopy part of g does not meet B_1 and so we see that $B_1 \subseteq f(B_0) \subseteq \Upsilon$.

In summary, we have $B_2 \subseteq B_1 \subseteq \Upsilon$. After subdividing, we can suppose that any cell of Υ meeting B_2 is disjoint from the spherical boundary, $S(B_1)$, of B_1. Let $\mathcal{P}$ be the set of ν-cells of Υ meeting B_2 in their relative interiors. Each of these is a

ν-block, and by the same dimension argument as in the proof of Lemma 4.2, we have $B_2 \subseteq \bigcup \mathcal{P}$. We claim that $\bigcup \mathcal{P} \subseteq \Phi$.

In fact suppose that $P \in \mathcal{P}$. We apply Lemma 4.4 with $X = \Upsilon$, $B = B_1$. Since Υ is a complex of dimension ν, we have $P \setminus S(P)$ open in Υ. Also, $P \setminus S(B_1) = \varnothing$, and by assumption $B_2 \cap P \setminus S(P) \subseteq B_1 \cap P \setminus S(P)$ is nonempty. It follows that $P \subseteq B_1$, so in particular, $P \subseteq \Phi$.

Since B_2 can be chosen arbitrarily, we see that any compact subset of Φ is contained in a finite union of ν-blocks contained in Φ, and so Φ is cubulated as required. □

Remark. In fact, the argument shows that if $B \subseteq M$ is homeomorphic to a closed ν-ball, and $K \subseteq B \setminus \partial B$ is a compact subset of the relative interior, then there is a compact cubulated set, Υ, with $K \subseteq \Upsilon \subseteq B$.

Combining Proposition 4.3 and Lemma 3.8, we get:

Proposition 4.6. *Suppose that M is a complete median metric space, and that $\Phi \subseteq M$ is a closed subset and that there is a homeomorphism $f : \mathbb{R}^\nu \to \Phi$ with the following property. For each codimension-1 coordinate plane, $H \subseteq \mathbb{R}^\nu$, there is a closed subset, $\Psi \subseteq M$, homeomorphic to $\mathbb{R}^\nu$ such that $f(H) = \Phi \cap \Psi$. Then Φ is a flat, and f is a median isomorphism.*

Note that the hypotheses on Φ only depend on the topological structure of M.

We conclude, in particular, that Φ is isometric to $\mathbb{R}^\nu$ with the l^1 metric.

This is all we will need for our discussion of the marking graph. We also include the following results which will be relevant to applications elsewhere; see [Bowditch 2015; 2016a].

Definition. We say that an $\mathbb{R}$-tree is *furry* if every point has valence at least 3.

Proposition 4.7. *Suppose that M is a complete median metric space of rank ν, that D is a direct product of ν furry $\mathbb{R}$-trees, and that $f : D \to M$ is a continuous injective map with closed image. Then f is a median homomorphism. Moreover, $f(D)$ is convex.*

Proof. By a *product flat* in D we mean a direct product of bi-infinite geodesics in each of the factors. If every point in each factor has valence at least 4 (as in the cases of genuine interest) then we see that every product flat Φ satisfies the hypotheses of Proposition 4.6, and so $f|\Phi$ is a median homomorphism. Now any two points, a, b lie in some such product flat Φ and $[a, b] \subseteq \Phi$. Thus, if $c \in [a, b]$, then $fc \in [fa, fb]$, and it follows that f is a median homomorphism on all of D.

If we allow for vertices of valence 3, then we just note that any codimension-1 coordinate plane in Φ is the intersection of three product flats, hence cubulated. We can then apply Lemma 3.8 directly, to see that f is a median homomorphism on Φ, hence, as above, everywhere. □

We remark that Proposition 4.7 applies in particular if M is also a product of ν $\mathbb{R}$-trees. It follows that f splits as a direct product of embeddings, up to permutation of the factors. Some further discussion of this, with applications, can be found in [Bowditch 2016b].

Definition. A *tree product*, T, in M is a convex subset median isomorphic to a direct product of ν nontrivial rank-1 median algebras. It is *maximal* if it is not contained in any strictly larger tree product.

Note that T is an l^1-product of $\mathbb{R}$-trees. It is easily seen that the closure of a tree product is a tree product, and so any maximal tree product is closed.

Note that in the above terminology, any closed subset of M homeomorphic to a direct product of ν furry $\mathbb{R}$-trees for $\nu \geq 2$ is a tree product by Proposition 4.7.

For applications elsewhere, in particular in [Bowditch 2016a], we note that we can relax the "furriness" condition somewhat.

Definition. An $\mathbb{R}$-tree is *almost furry* if it is infinite, and no point has valence equal to 2.

In this case, by removing the extreme (valence-1) points, we obtain the maximal furry subtree. The following was given as Proposition 1.3 in the introduction.

Proposition 4.8. *Suppose that M is a complete median metric space of rank ν, that D is a direct product of ν almost furry $\mathbb{R}$-trees, and that $f : D \to M$ is a continuous injective map with closed image. Then f is a median homomorphism. Moreover, $f(D)$ is convex.*

Proof. We can apply the arguments to the maximal subset which is a product of furry trees, and then take its closure. We have already observed that the key statements, in particular Lemma 3.8 and Proposition 4.3, have local versions which can be applied to this case. □

5. Ultraproducts

In this section, we give some general background to the theory of ultraproducts and asymptotic cones. The notion of an asymptotic cone was introduced in [van den Dries and Wilkie 1984]; see also [Gromov 1993]. The idea behind this is to keep rescaling the metric so that points move closer and closer together, and then pass to an "ultralimit" of the resulting spaces. (Here, the term "ultralimit" is used in the sense of [Gromov 1993], rather than in the usual sense of model theory.) We then factor out "infinitesimals" to give what we call here an "extended asymptotic cone". If we also throw away the "unlimited" parts (beyond infinity), we get the usual asymptotic cone. In principle, this may depend on the choice of rescaling factors and (if the continuum hypothesis fails) on the choice of ultrafilter, but such ambiguity will not matter to us here.

Let $\mathcal{Z}$ be a countable set equipped with a nonprincipal ultrafilter. We can think of this as a finitely additive measure on $\mathcal{Z}$, taking values in $\{0, 1\}$, such that $\mathcal{Z}$ itself has measure 1, and any finite subset of $\mathcal{Z}$ has measure 0. If a predicate, $P(\zeta)$, depends on $\zeta \in \mathcal{Z}$, we say that P holds *almost always* if the set of ζ for which it holds has measure 1.

We refer to a sequence of objects indexed by $\mathcal{Z}$ as a *$\mathcal{Z}$-sequence*. Typically, we will use the notation $\vec{X} = (X_\zeta)_\zeta$ for such a sequence. If these are all sets, we write $\prod \vec{X} = \prod_\zeta X_\zeta$ for their product. Given $\vec{x}, \vec{y} \in \prod \vec{X}$, we write $\vec{x} \approx \vec{y}$ to mean that $x_\zeta = y_\zeta$ almost always. Thus, $\approx$ is an equivalence relation on $\prod \vec{X}$, and we write $\mathcal{U}\vec{X} = \prod \vec{X}/{\approx}$ for the quotient.

Definition. We refer to $\mathcal{U}\vec{X}$ as the *ultraproduct* of the $\mathcal{Z}$-sequence $\vec{X}$.

Note that we only need to have x_ζ defined almost always to determine an element of $\mathcal{U}\vec{X}$. We write $\boldsymbol{x} = [\vec{x}]$ for this element.

We write $\mathcal{P}(\vec{X})$ for the $\mathcal{Z}$-sequence $(\mathcal{P}(X_\zeta))_\zeta$, where $\mathcal{P}$ denotes power set. There is a natural map $\mathcal{U}\mathcal{P}(\vec{X}) \to \mathcal{P}(\mathcal{U}\vec{X})$, defined by sending $\vec{Y}$ to the set of $\boldsymbol{x} = [\vec{x}] \in \mathcal{U}\vec{X}$ such that $x_\zeta \in Y_\zeta$ almost always. We can identify the image of this map with $\mathcal{U}\vec{Y}$. Note that we can define unions and intersections in $\mathcal{P}(\vec{X})$ (by taking unions and intersections on each ζ-coordinate). These operations are respected by the above map.

Given two $\mathcal{Z}$-sequences of sets, $\vec{X}$ and $\vec{Y}$, we can form the direct product $\vec{X} \times \vec{Y}$ as $(X_\zeta \times Y_\zeta)_\zeta$, and we see that $\mathcal{U}(\vec{X} \times \vec{Y})$ is naturally identified with $\mathcal{U}\vec{X} \times \mathcal{U}\vec{Y}$. A $\mathcal{Z}$-sequence of relations on $X_\zeta \times Y_\zeta$ give rise to a relation on $\mathcal{U}\vec{X} \times \mathcal{U}\vec{Y}$ via the map from $\mathcal{U}\mathcal{P}(\vec{X} \times \vec{Y})$ to $\mathcal{P}(\mathcal{U}\vec{X} \times \mathcal{U}\vec{Y})$. In other words, $\boldsymbol{x}$ is related to $\boldsymbol{y}$ if x_ζ is almost always related to y_ζ. A particular case is when relation on $X_\zeta \times Y_\zeta$ is almost always the graph of a function. In fact, the following is a simple exercise:

Lemma 5.1. *Given any $\mathcal{Z}$-sequence of functions, $f_\zeta : X_\zeta \to Y_\zeta$, there is a unique function $\mathcal{U}f : \mathcal{U}\vec{X} \to \mathcal{U}\vec{Y}$, such that $\boldsymbol{y} = \mathcal{U}f(\boldsymbol{x})$ if and only if $y_\zeta = f_\zeta(x_\zeta)$ almost always.*

We also note that the discussion of relations also applies to finite products of sets, and so to n-ary relations and n-ary operations for any finite n. For example, if $\vec{\Gamma}$ is a sequence of groups, then $\mathcal{U}\vec{\Gamma}$ has the structure of a group. If each Γ_ζ acts on a set X_ζ, then $\mathcal{U}\vec{\Gamma}$ acts on $\mathcal{U}\vec{X}$.

Suppose that $X_\zeta = X$ is constant. In this case, we write $\mathcal{U}X = \mathcal{U}\vec{X}$.

Definition. We refer to $\mathcal{U}X$ as the *ultrapower* of the set X.

There is a natural injection X into $\mathcal{U}X$ obtained by taking constant sequences. We refer to the image of this map as the *standard part* of $\mathcal{U}X$. We usually identify X with the standard part of $\mathcal{U}X$. If X is finite, then $\mathcal{U}X$ is equal to its standard part.

Note that the ultrapower, $\mathcal{U}\mathbb{R}$, of the real numbers is an ordered field. We say that $\boldsymbol{x} \in \mathcal{U}\mathbb{R}$ is *limited* if $|\boldsymbol{x}| \leq \boldsymbol{y}$ for some $\boldsymbol{y} \in \mathbb{R} \subseteq \mathcal{U}\mathbb{R}$ (where $|\boldsymbol{x}| = \max\{\boldsymbol{x}, -\boldsymbol{x}\}$).

Otherwise it is *unlimited*. We say that $\boldsymbol{x}$ is *infinitesimal* if $|\boldsymbol{x}| \leq \boldsymbol{y}$ for all positive standard $\boldsymbol{y}$. Note that 0 is the only standard infinitesimal, and that the nonzero infinitesimals are exactly the reciprocals of unlimited numbers.

There is a well defined map $\mathrm{st} : \mathcal{U}\mathbb{R} \to [-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\}$ such that $\mathrm{st}(\boldsymbol{x}) = \infty$ if $\boldsymbol{x}$ is positive unlimited, $\mathrm{st}(\boldsymbol{x}) = -\infty$ if $\boldsymbol{x}$ is negative unlimited, and $\boldsymbol{x} - \mathrm{st}(\boldsymbol{x})$ is infinitesimal if x is limited. We refer to $\mathrm{st}(\boldsymbol{x})$ as the *standard part* of $\boldsymbol{x}$. We will usually restrict attention to nonnegative numbers, so we get a map $\mathrm{st} : \mathcal{U}[0, \infty) \to [0, \infty]$. If $(x_\zeta)_\zeta$ is a $\mathcal{Z}$-sequence of real numbers, we write $x_\zeta \to x \in \mathbb{R} \cup \{\infty\}$ to mean that $x = \mathrm{st}(\boldsymbol{x})$. (This is the same as taking limits in $\mathbb{R}$ with respect to the ultrafilter.)

In the case of the natural number, there are no infinitesimals, and $\mathbb{N}$ is an initial segment of $\mathcal{U}\mathbb{N}$. We get a map $\mathrm{st} : \mathcal{U}\mathbb{N} \to \mathbb{N} \cup \{\infty\}$ which is the identity on $\mathbb{N}$.

Given any set M we define an *nonstandard metric* on M to be a metric with values in $\mathcal{U}\mathbb{R}$. In other words, it is a map $M^2 \to \mathcal{U}[0, \infty)$ satisfying the same axioms as a metric, except with $\mathbb{R}$ replaced by $\mathcal{U}\mathbb{R}$. Note that, if σ is a nonstandard metric, the composition $\hat{\sigma} = \mathrm{st} \circ \sigma : M^2 \to [0, \infty]$ is an idealised pseudometric on M. Here, we use the term *idealised* to mean that we allowing points to be an infinite distance apart. As with usual pseudometric spaces, we can take the Hausdorffification, $\hat{M}$, of M. In other words, given $x, y \in M$, we write $x \simeq y$ to mean that $\hat{\sigma}(x, y) = 0$. Thus $\simeq$ is an equivalence relation on M, and we set $\hat{M} = M/\simeq$. The induced map, $\hat{\sigma} : \hat{M}^2 \to [0, \infty]$ is an idealised metric on $\hat{M}$. Note that the relation on $\hat{M}$ given by deeming x to be equivalent to y if $\hat{\sigma}(x, y) < \infty$ is an equivalence relation.

Definition. A *component* of an idealised metric space, $\hat{M}$, is an equivalence class under the above relation.

Note that components are both open and closed in the topology induced on $\hat{M}$. Also any component is a metric space in the usual sense. Note that we can speak about an extended metric space as being *complete*; that is, all its components are complete metric spaces. (We will see that, in the cases of interest in this paper, this notion "component" will coincide with the usual notion of a connected component).

Suppose that $((X_\zeta, \sigma_\zeta))_\zeta$ is a $\mathcal{Z}$-sequence of metric spaces. This gives rise to a nonstandard metric, $\mathcal{U}\sigma$, on $\mathcal{U}\vec{X}$, and hence to an idealised pseudometric, $\hat{\sigma}$, on $\mathcal{U}\vec{X}$. Let $\hat{X}$ be the Hausdorffification, with idealised metric $\hat{\sigma} : \hat{X}^2 \to [0, \infty]$.

If $x_\zeta \in X_\zeta$, we write $x_\zeta \to x$ to mean that $x \in \hat{X}$ is the image of the sequence $\boldsymbol{x}$ under the natural maps. We think of x as the limit of the x_ζ. By construction, every sequence has a unique limit.

For the following lemma, we use the fact that $\mathcal{Z}$ is countable to find a $\mathcal{Z}$-sequence $(n_\zeta)_\zeta$ in $\mathbb{N}$ with $n_\zeta \to \infty$. For example let, $n : \mathcal{Z} \to \mathbb{N}$ be any injective map. (This is true of a broader class of cardinals than $\aleph_0$, though we won't pursue that issue here — we will have no need of any uncountable indexing sets.)

Lemma 5.2. *$(\hat{X}, \hat{\sigma})$ is complete.*

Proof. Let $(x^i)_{i\in\mathbb{N}}$ be a Cauchy sequence in $\hat{X}$. It is enough to show that $(x^i)_i$ has a convergent subsequence. We can suppose that $\hat{\sigma}(x^i, x^{i+1}) \le 1/2^{i+1}$ for all i. Given $i \in \mathbb{N}$, let $(x^i_\zeta)_\zeta$ be some representative of x^i in $\vec{X} = (X_\zeta)_\zeta$. Let $\mathcal{Z}_0(j) = \{\zeta \in \mathcal{Z} \mid \sigma_\zeta(x^j_\zeta, x^{j+1}_\zeta) \le 1/2^j\}$, let $\mathcal{Z}(i) = \bigcap_{j\le i} \mathcal{Z}_0(j)$ and $\mathcal{Z}(\infty) = \bigcap_{j=0}^\infty \mathcal{Z}_0(j)$. Given $\zeta \in \mathcal{Z} \setminus \mathcal{Z}(\infty)$, let $i(\zeta) = \max\{i \mid \zeta \in \mathcal{Z}(i)\}$, and let $y_\zeta = x^{i(\zeta)}_\zeta$. Note that if $\zeta \in \mathcal{Z}(i)$, then $\sigma_\zeta(y_\zeta, x^i_\zeta) \le \sum_{j\ge i}(1/2^i) \le 2/2^i$. We distinguish two cases.

If $\mathcal{Z}(\infty)$ has measure 0, then y_ζ is defined almost always. Let y be the image of $(y_\zeta)_\zeta$ in $\hat{X}$. Now $\sigma_\zeta(y_\zeta, x^i_\zeta) \le 2/2^i$ almost always, and so $\hat{\sigma}(y, x^i) \le 2/2^i$, showing that x^i converges to y.

If $\mathcal{Z}(\infty)$ has measure 1, we set $y_\zeta = x^{n_\zeta}_\zeta$, where $n_\zeta \to \infty$, and argue as before. □

Suppose that $A_\zeta \subseteq X_\zeta$ (almost always). As discussed earlier, this gives rise to a subset of $\mathcal{U}\vec{X}$ which can be identified with $\mathcal{U}\vec{A}$. We denote its image in $\hat{X}$ by $\hat{A}$. In fact, restricting the metrics, $(\hat{A}, \hat{\sigma})$ is the limit of the subspaces (A_ζ, σ_ζ) constructed intrinsically. Note that $x \in \hat{A}$ if and only if $\sigma_\zeta(x_\zeta, A_\zeta) \to 0$ (where we are taking limits with respect to the ultrafilter on $\mathcal{Z}$). We also note that $\hat{A}$ is closed in the induced topology on $\hat{X}$. This can be seen by a similar argument to Lemma 5.2, or simply by noting that $\hat{A}$ is complete in the induced metric. Note that $\hat{\mathbb{R}}$ is an ordered abelian group, which we refer to as the *extended reals*. (In this paper, this will usually be denoted instead by $\mathbb{R}^*$, for the reasons given below.)

Suppose that $f_\zeta : X_\zeta \to Y_\zeta$ is a $\mathcal{Z}$-sequence of maps between the metric spaces (X_ζ, σ_ζ) and (Y_ζ, σ'_ζ). We have a map, $\mathcal{U}f : \mathcal{U}\vec{X} \to \mathcal{U}\vec{Y}$ given by Section 1. Suppose there is a constant, $k \in [0, \infty)$, and a $\mathcal{Z}$-sequence, $(h_\zeta)_\zeta$, in $[0, \infty)$ with $h_\zeta \to 0$, such that for almost all ζ and all $x, y \in X_\zeta$ we have $\sigma'_\zeta(f_\zeta(x), f_\zeta(y)) \le k\sigma_\zeta(x, y) + h_\zeta$. Then, $\mathcal{U}f$ induces a k-Lipschitz map $\hat{f} : \hat{X} \to \hat{Y}$. (The graph of $\hat{f}$ is the limit of the graphs of the f_ζ, taking the l^1 metrics on $X_\zeta \times Y_\zeta$.) The image $\hat{f}(\hat{Y})$ is the limit of the images, $f_\zeta(X_\zeta)$, in the sense of the previous paragraph.

Suppose that $((X_\zeta, \mathcal{Z}))_\zeta$ is a $\mathcal{Z}$-sequence of geodesic metric spaces. Then the components of $(\hat{X}, \hat{\sigma})$ are precisely the connected components, and each such component is a geodesic space. (This can be seen by applying the previous paragraph to geodesics, thought of as uniformly Lipschitz maps of a compact real interval into the spaces X_ζ.)

Suppose that $(X_\zeta, \rho_\zeta) = (X, \rho)$ a constant sequence. In this case, we get a natural injective map of (X, ρ) into the limit $(\hat{X}, \hat{\rho})$, which is an isometry onto its range. The closure of this range in $\hat{X}$ is just the metric completion of X.

More interestingly, we can take a positive infinitesimal, $\boldsymbol{t} \in \mathcal{U}\mathbb{R}$, and set $\sigma_\zeta = t_\zeta \rho$ to be the rescaled pseudometric. In this case, we write $(X^*, \rho^*) = (\hat{X}, \hat{\sigma})$ for the limiting space. Note that this is the same as taking the rescaled metric space $(\hat{X}, \boldsymbol{t}\hat{\rho})$ and passing to its Hausdorffification.

Definition. We refer to (X^*, ρ^*) as the *extended asymptotic cone* of X with respect to $\boldsymbol{t}$.

Note that X^* has a preferred basepoint, namely that given by any constant sequence in $\vec{X}$. This, in turn, determines a preferred component, X^∞, of X^*, namely that containing this basepoint.

Definition. We refer to X^∞ as the *asymptotic cone* of X with respect to $\boldsymbol{t}$.

By Lemma 5.2, the asymptotic cone is always complete. If X is a geodesic space, so is X^∞.

One can generalise the above to a $\mathcal{Z}$-sequence of metric spaces, (X_ζ, ρ_ζ), rescaled by an infinitesimal $\boldsymbol{t}$, to give an extended asymptotic cone, (X^*, ρ^*). In this case, one needs a sequence of basepoints, $e_\zeta \in X_\zeta$ to determine a basepoint and base component of X^*.

Definition. We say that a $\mathcal{Z}$-sequence of maps, $f_\zeta : X_\zeta \to Y_\zeta$ between metric spaces are *uniformly coarsely Lipschitz* if there are constants, $k, h \geq 0$, such that for almost all $\zeta \in \mathcal{Z}$ and all $x, y \in X_\zeta$, we have $\sigma'_\zeta(f_\zeta x, f_\zeta y) \leq k\sigma_\zeta(x, y) + h$. They are *uniform quasi-isometric embeddings* if also $\sigma_\zeta(x, y) \leq k\sigma'_\zeta(f_\zeta x, f_\zeta y) + h$. They are *uniform quasi-isometries* if also $Y_\zeta = N(f_\zeta(X_\zeta); h)$.

Lemma 5.3. *A $\mathcal{Z}$-sequence of uniformly coarsely Lipschitz maps, $f_\zeta : X_\zeta \to Y_\zeta$, induces a Lipschitz map, $f^* : X^* \to Y^*$, which restricts to a map $f^\infty : X^\infty \to Y^\infty$. If the maps f_ζ are uniform quasi-isometric embeddings, then f^* and f^∞ are bi-Lipschitz onto their range. If they are quasi-isometries, then f^* and f^∞ are bi-Lipschitz homeomorphisms.*

Proof. By Lemma 5.1, we have a map $\mathcal{U}f : \mathcal{U}\boldsymbol{X} \to \mathcal{U}\boldsymbol{Y}$. This descends to a map $f^* : X^* \to Y^*$ (since $t_\zeta\sigma_\zeta(x_\zeta, y_\zeta) \to 0$ implies $t_\zeta\sigma'_\zeta(f_\zeta(x_\zeta), f_\zeta(y_\zeta)) \to 0$). The fact that f^* and its restriction f^∞ are (bi-)Lipschitz follows since the $ht_\zeta \to 0$, so the additive constant disappears in the limit. □

In particular, quasi-isometric spaces have bi-Lipschitz equivalent asymptotic cones (for the same scaling sequence).

An example of the above construction is given by a sequence, $\vec{G} = (G_\zeta)_\zeta$ of graphs. Let $V_\zeta = V(G_\zeta)$ be the vertex sets. The adjacency relations on the V_ζ determine an adjacency relation on $\mathcal{U}\vec{V}$, so as to give it the structure of the vertex set $V(\mathcal{U}\vec{G})$ of a graph $\mathcal{U}\vec{G}$. If each G_ζ is connected, the combinatorial distance functions on V_ζ give us a limiting nonstandard metric and hence an idealised metric on $\mathcal{U}\vec{V}$, with values in $\mathbb{N} \cup \{\infty\}$. This is the same as the combinatorial idealised metric given by $\mathcal{U}V = V(\mathcal{U}\vec{G})$. In particular, the components are again the connected components. (Note that we lose some information in the standardisation process, since different pairs of components might be at different unlimited distances apart.)

Suppose that $\vec{\Gamma} = (\Gamma_\zeta)_\zeta$ is a $\mathcal{Z}$-sequence of groups. Then $\mathcal{U}\vec{\Gamma}$ is also a group. If each Γ_ζ acts on a set X_ζ, then $\mathcal{U}\vec{\Gamma}$ acts on $\mathcal{U}\vec{X}$. If Γ_ζ acts by isometry in some metric space, then so does $\mathcal{U}\vec{\Gamma}$. If Γ and X are fixed, then any two points of $X \subseteq \mathcal{U}\vec{X}$ in the same $\mathcal{U}\Gamma$-orbit also lie in the same Γ-orbit (since if $y = gx$ for some $g \in \mathcal{U}\Gamma$, then $y = g_\zeta x$ for almost all g_ζ, and so certainly for some g_ζ).

If Γ is a fixed group acting on a metric space, X, we get an induced action of $\mathcal{U}\Gamma$ on the extended asymptotic cone, X^* (with respect to any infinitesimal $\boldsymbol{t}$). Note that we can identify Γ as a normal subgroup of $\mathcal{U}\Gamma$. In fact, we have normal subgroups, $\Gamma \lhd \mathcal{U}^1\Gamma$, $\mathcal{U}^1\Gamma \lhd \mathcal{U}^0\Gamma$ and $\mathcal{U}^0\Gamma \lhd \Gamma_\infty$ of Γ_∞, where $\mathcal{U}^1\Gamma$ is the stabiliser of the basepoint of X^*, and $\mathcal{U}^0\Gamma$ is the setwise stabiliser of the asymptotic cone, X^∞. Note that $\mathcal{U}^1\Gamma$ and $\mathcal{U}^0\Gamma$ may depend on $\boldsymbol{t}$.

If the action of Γ on X is cobounded (i.e., X is a bounded neighbourhood of some, hence any, Γ-orbit), then the actions of $\mathcal{U}\Gamma$ on X^* and of $\mathcal{U}^0\Gamma$ on X^∞ are transitive. In particular, X^* and X^∞ are homogeneous (extended) metric spaces.

Note that a special case of this construction is $\mathbb{R}^*$, which is always isomorphic to the extended reals, $\hat{\mathbb{R}}$. If X is a Gromov hyperbolic space, then X^* is an $\mathbb{R}^*$-tree, and X^∞ is an $\mathbb{R}$-tree. Of course, this also applies to the asymptotic cone of a sequence of uniformly hyperbolic spaces.

Terminology. To briefly summarise our terminology, we use "nonstandard" to refer to ultralimits, "extended" to refer to the standard part of a nonstandard number (quotienting out by infinitesimals), and "idealised" to mean we are adjoining $\pm\infty$. In this way, we have the extended reals $\mathbb{R}^*$ as a subset (or quotient) of the nonstandard reals $\mathcal{U}\mathbb{R}$. We can view the idealised reals, $[-\infty, \infty]$, as a quotient of $\mathbb{R}^*$.

6. Coarse median spaces

Coarse median spaces were defined in [Bowditch 2013]. The main point here is that they give a means of talking about (quasi)cubes or (quasi)flats in a geodesic space. Following the construction of [Behrstock and Minsky 2011], this is applicable to the mapping class group, as shown in [Bowditch 2013]. It also applies to Teichmüller space in either the Teichmüller metric, [Bowditch 2016a] or the Weil–Petersson metric [Bowditch 2015]. We remark that another class of space which encompasses these cases, and which implies coarse median, is described in [Behrstock, Hagen and Sisto 2015; 2017].

Before continuing, we introduce the following general conventions.

Conventions. Given two points, x, y, in a metric space, and $r \geq 0$, we will write $x \sim_r y$ to mean that the distance between them is at most r. We will often simply write $x \sim y$, and behave as though this relation were transitive. Here is understood that, at any given stage, the bound r depends only on the constants introduced at the beginning of an argument. It can be explicitly determined by following through the steps of the argument, though we will not usually explicitly estimate it.

Similarly, given two functions f, g, we will write $f \sim g$ to mean that $f(x) \sim g(x)$ for all x in the domain.

We are often only interested in maps defined up to bounded distance. For a graph it would therefore be enough to specify a map on the set of vertices. When referring to a finite product of metric spaces, we can always take the l^1 metric. For a finite product of graphs, we can always restrict to the 1-skeleton of the product cube complex. In any case, we will only be interested in the product metric defined up to bi-Lipschitz equivalence.

We will sometimes adopt a similar convention for linear bounds. Given $\lambda \geq 1$ and $r \geq 0$, we write $x \asymp_{\lambda,r} y$ to mean that $\lambda^{-1}(x - r) \leq y \leq \lambda x + r$. Again, we usually omit λ, r from the notation, and write $x \asymp y$.

When we come to discuss marking graphs, the constants implicit in the notation $\sim$ and $\asymp$ will ultimately depend only on the complexity, $\xi(\Sigma)$, of our surface, Σ, as defined in Section 7. We will make this explicit at the relevant points.

Let $(\mathcal{M}, \rho)$ be a geodesic metric space.

Definition. We say that a ternary operation, $\mu : \mathcal{M}^3 \to \mathcal{M}$, is a "coarse median" if it satisfies the following:

(C1) There are constants, k, $h(0)$, such that for all $a, b, c, a', b', c' \in \mathcal{M}$ we have $\rho(\mu(a, b, c), \mu(a', b', c')) \leq k(\rho(a, a') + \rho(b, b') + \rho(c, c')) + h(0)$, and

(C2) There is a function, $h : \mathbb{N} \to [0, \infty)$, with the following property. Suppose that $A \subseteq \mathcal{M}$ with $1 \leq |A| \leq p < \infty$. Then there is a median algebra, (Π, μ_Π) and maps $\pi : A \to \Pi$ and $\lambda : \Pi \to \mathcal{M}$ such that for all $x, y, z \in \Pi$ we have $\rho(\lambda\mu_\Pi(x, y, z), \mu(\lambda x, \lambda y, \lambda z)) \leq h(p)$ and for all $a \in A$, we have $\rho(a, \lambda\pi a) \leq h(p)$.

We say that $\mathcal{M}$ has *rank at most n* if we can always take Π to have rank at most n (as a median algebra). We say that M is *n-colourable* if we can always take Π to be n-colourable. We refer to $(\mathcal{M}, \rho, \mu)$ as a *coarse median space*. We refer to k, h as the *parameters* of $\mathcal{M}$.

From (C2) we can deduce that, if $a, b, c \in \mathcal{M}$, then $\mu(a, b, c)$, $\mu(b, a, c)$ and $\mu(b, c, a)$ are a bounded distance apart, and that $\rho(\mu(a, a, b), a)$ is bounded. Since we are only really interested in μ up to bounded distance, we can assume that μ is invariant under permutation of a, b, c and that $\mu(a, a, b) = a$.

Note that in (C2), we can always assume that $\Pi = \langle \pi A \rangle$ (in particular, that it is finite). Also, if we are not concerned about rank, we can always take Π to be the free median algebra on A, and π to be the inclusion of A in Π.

Note that a direct of product of coarse median spaces is also a coarse median space.

For future reference, we note:

Lemma 6.1. *Suppose $a, b, c \in \mathcal{M}$, and $r \geq 0$ with $\rho(\mu(a, b, c), c) \leq r$. Then $\rho(a, c) + \rho(c, b) \leq k_1 \rho(a, b) + k_2$, where k_1 and k_2 depend only on the parameters of $\mathcal{M}$.*

Proof. Using property (C1), we see that the maps $x \mapsto \mu(a, c, x)$ and $x \mapsto \mu(b, c, x)$ are coarsely Lipschitz, and so we get linear bounds on $\rho(a, c)$ and $\rho(b, c)$ in terms of $\rho(a, b)$. $\square$

With the conventions introduced earlier, this shows $\rho(a, b) \asymp \rho(a, c) + \rho(c, b)$. (where the implicit constants depend only on the parameters of $\mathcal{M}$).

Given two spaces X, Y, equipped with ternary operations μ_X and μ_Y, together with a metric ρ on Y, we say that a map $\phi : X \to Y$ is an *l-quasimorphism* if $\rho(\phi \mu_X(x, y, z), \mu_Y(\phi x, \phi y, \phi z)) \leq l$ for all $x, y, z \in X$. Typically, Y will be a coarse median space, and X will be either a median algebra or a coarse median space. (Note that the map λ featuring in (C2) is an $h(p)$-quasimorphism.)

Lemma 6.2. *Suppose that Π is a median algebra generated by a finite subset, $B \subseteq \Pi$. Suppose that $\lambda, \lambda' : \Pi \to \mathcal{M}$ are l-quasimorphisms with $\rho(\lambda b, \lambda' b) \leq l$ for all $b \in B$. Then, for all $x \in \Pi$, $\rho(\lambda x, \lambda' x)$ is bounded above by some linear function of l, depending only on the parameters of $\mathcal{M}$ and the cardinality of B.*

Proof. Define $B_i \subseteq \Pi$ inductively by $B_0 = B$ and $B_{i+1} = \mu(B_i^3)$. We see inductively that $\lambda | B_i$ and $\lambda' | B_i$ are a bounded distance apart, where the bound depends on i and is linear in l. Now $|\Pi| \leq q = 2^{2^p}$ where $p = |B|$, and so certainly, $\Pi = B_q$, and the result follows. $\square$

In particular, in clause (C2) of the definition, if we assume that $\Pi = \langle \pi A \rangle$, then the map λ is unique up to bounded distance depending only on the parameters and p.

The following will allow us to assume that quasimorphisms of cubes are in fact uniform quasimorphisms:

Lemma 6.3. *Given $n \in \mathbb{N}$, there are constants k_0, h_0 and h_1 depending only on n and the parameters of $\mathcal{M}$ such that the following holds. Suppose that $Q = \{-1, 1\}^n$ and that $\psi : Q \to \mathcal{M}$ is an l-quasimorphism for some $l \geq 0$. Then there is an h_0-quasimorphism, $\phi : Q \to \mathcal{M}$, with $\rho(\phi x, \psi x) \leq k_0 l + h_1$ for all $x \in Q$.*

Proof. Let Π be the free median algebra on the set Q, and let $\theta : \Pi \to Q$ be the unique median homomorphism extending the identity on Q (thought of as a map from a set to a median algebra). Now there is a median monomorphism, $\omega : Q \to \Pi$ with $\theta \circ \omega$ the identity on Q. (To see this, we can think of Π as the vertex set of a finite CAT(0) cube complex. Every pair of intrinsic faces of $Q \subseteq \Pi$ are separated by some hyperplane of Π, and these must all intersect in some n-cell of Π. Each element, $x \in Q$, determines a unique vertex $\omega(x)$ of this n-cell. This gives us a homomorphism $\omega : Q \to \Pi$, with $\omega(Q)$ equal to the vertex set of the n-cell. Note that ω is not canonically determined: it might depend on the choice of cell.)

Now apply (C2) to $\psi(Q) \subseteq \mathcal{M}$ to give an $h(2^n)$-quasimorphism $\lambda : \Pi \to \mathcal{M}$ with $\lambda|Q \sim_{h(2^n)} \psi$. Let $\phi = \lambda \circ \omega : Q \to \mathcal{M}$. This is an h_0-quasimorphism, where $h_0 = h(2^n)$.

Let $\lambda' = \lambda \circ \theta : \Pi \to \mathcal{M}$. Thus λ' is a l-quasimorphism, and $\lambda'|Q = \psi = \lambda|Q$. By Lemma 6.2, we have $\rho(\lambda x, \lambda' x) \leq k_0 l + h_2$ for all $x \in \Pi$, where k_0, h_2 depend only on the parameters of $\mathcal{M}$. But $\lambda' \circ \omega|Q = \lambda \circ \theta \circ \omega|Q = \lambda|Q \sim_{h(2^n)} \psi$, and so we see that $\rho(\phi x, \psi x) \leq k_0 l + h_1$ for all $x \in Q$, where $h_1 = h_2 + h(2^n)$. □

The following two lemmas will be used to establish that statements that hold in a median algebra hold up to bounded distance in a coarse median space.

Lemma 6.4. *Suppose that $(\mathcal{M}, \rho, \mu)$ is a coarse median space. Suppose that Π is a finite median algebra with $|\Pi| \leq q < \infty$, and that $\lambda : \Pi \to \mathcal{M}$ is an l-quasimorphism. Given $t \geq 0$, there is a finite median algebra Π', a map $\lambda' : \Pi' \to \mathcal{M}$ and an epimorphism $\theta : \Pi \to \Pi'$ such that for all distinct $x, y \in \Pi'$, $\rho(\lambda' x, \lambda' y) > t$, and for all $z \in \Pi$, $\rho(\lambda z, \lambda' \theta z) \leq s$, where s depends only on q, h, t and the parameters of $\mathcal{M}$.*

Proof. Define a relation, $\approx$, on Π, by setting $x \approx y$ if $\rho(\lambda x, \lambda y) \leq t$. Let $\simeq$ be the smallest equivalence relation on Π containing $\approx$ with the property that whenever $x, y, z, w \in \Pi$ with $z \simeq w$, we have $\mu_\Pi(x, y, z) \simeq \mu_\Pi(x, y, w)$. Let $\Pi' = \Pi/\simeq$ be the quotient median algebra as defined at the end of Section 2, and let $\theta : \Pi \to \Pi'$ be the quotient map. Define $\lambda' : \Pi' \to \mathcal{M}$ by setting $\lambda'(x)$ to be the λ-image of any representative of the $\simeq$-class of x in Π. Since $\simeq$ includes $\approx$, we see that $\rho(\lambda' x, \lambda' y) > t$ for all distinct $x, y \in \Pi'$.

We claim that if $x, y \in \Pi$, with $x \simeq y$, then $\rho(\theta\lambda x, \theta\lambda y)$ is bounded above in terms of q, h, t and the parameters of $\mathcal{M}$. To see this, note that $\simeq$ can be constructed from $\approx$ by iterating two operations. We start with $\approx$. Whenever $z \approx w$, then we set $\mu_\Pi(x, y, z)$ related to $\mu_\Pi(x, y, w)$ for all $x, y \in \Pi$. Also, if $a \approx b$ and $b \approx c$, then we set a to be related to c. We continue again with the relation thus defined. After at most q steps, this process stabilises on the relation $\simeq$. From the fact that λ is a quasimorphism, and from property (C1) for $\mathcal{M}$, we see that at each stage the maximal distance between the λ-images of related elements of Π can increase by at most a linear function which depends only on l and the parameters of $\mathcal{M}$. This now proves the claim.

Suppose that $z \in \Pi$. By construction, $\lambda' \theta z = \lambda w$ for some $w \simeq z$. By the above, $\rho(\lambda z, \lambda' \theta z) = \rho(\lambda z, \lambda w)$ is bounded as required. □

Note that λ' is itself an l'-quasimorphism, where l' depends only on q, h, t and the parameters of $\mathcal{M}$. This enables us to give a refinement of (C2) as follows:

Corollary 6.5. *Suppose that $(\mathcal{M}, \rho, \mu)$ is a coarse median space, and $t \geq 0$. Then there is a function, $h_t : \mathbb{N} \to [0, \infty)$ with the following property. Suppose that $A \subseteq \mathcal{M}$ with $1 \leq |A| \leq p < \infty$. Then there is a finite median algebra, (Π, μ_Π), and maps $\pi : A \to \Pi$ and $\lambda : \Pi \to \mathcal{M}$ such that λ is a $h_t(p)$-quasimorphism with $\rho(\lambda x, \lambda y) > t$ for all distinct $x, y \in \Pi$, and such that $\rho(a, \lambda\pi a) \leq h_t(p)$ for all $a \in A$.*

Proof. Start with Π, π, λ as given by (C2) for $\mathcal{M}$ (so that $\lambda : \Pi \to \mathcal{M}$ is an $h(p)$-quasimorphism, with $h(p)$ independent of t). We can assume that $|\Pi| \leq 2^{2^p}$. We now apply Lemma 6.4 to give Π', λ' and $\theta : \Pi \to \Pi'$. Now replace Π by Π', π by $\theta \circ \pi$, and λ by λ'. ☐

By an "identity" in a median algebra, we mean an expression equating two terms featuring only the median operation. We refer to it as a "tautological identity" if it holds, whatever the arguments in any median algebra, M. (For example, we have the tautological identity: $\mu(a, b, \mu(a, b, c)) = \mu(a, b, c)$, for all $a, b, c \in M$.) We remark that an identity can easily be verified algorithmically: it is sufficient to check it for all possible assignments of the arguments in the two-point median algebra $\{-1, 1\}$. We make the following general observation.

General Principle. Any tautological median identity holds up to bounded distance in any coarse median space, $\mathcal{M}$.

More formally, this says that if P and Q are formulae defining $P(a_1, \dots, a_n)$ and $Q(a_1, \dots, a_n)$ in terms of μ, and the identity $P(a_1, \dots, a_n) = Q(a_1, \dots, a_n)$ holds for any $a_1, \dots, a_n \in M$ in any median algebra (M, μ) then it follows that $\rho(P(a_1, \dots, a_n), Q(a_1, \dots, a_n))$ is bounded for any $a_1, \dots, a_n \in \mathcal{M}$ in any coarse median space $(\mathcal{M}, \mu, \rho)$. The bound only depends on (the complexity of) the formulae P, Q and the parameters of $\mathcal{M}$.

For example, for all $a, b, c \in \mathcal{M}$, $\rho(\mu(a, b, \mu(a, b, c)), \mu(a, b, c))$ is bounded above by a constant depending only on the parameters of $\mathcal{M}$ for all $a, b, c \in \mathcal{M}$.

To prove this principle, let $A \subseteq \mathcal{M}$ be the set of elements occurring as arguments, and let $\pi : A \to \Pi$ and $\lambda : \Pi \to \mathcal{M}$ be as given by (C2) of the hypotheses. Now apply either side of the identity to the π-images in Π to give an element $x \in \Pi$ (by assumption, this will be the same element for either side). We can also apply each side of the same identity to the elements of A, using the median structure, μ, on $\mathcal{M}$. In this way, we get two elements of $\mathcal{M}$. Using (C1) and (C2) directly, we see that these are both a bounded distance from λx, and so, a bounded distance from each other. The claim follows.

A more general statement holds for conditional identities. Suppose that some finite set of identities (the "input identities") imply another identity (the "derived identity") in any median algebra. (For example, $d \in [a,b] \cap [b,c] \cap [c,a]$ implies $d = \mu(a,b,c)$.) We have the following generalisation.

General Principle. Given a finite set of input identities, and a derived identity, if we suppose that the input identities hold up to bounded distance for a particular set of elements in a coarse median space, $\mathcal{M}$, then the derived identity also holds up to bounded distance for this set of elements.

(So, for example, if $a,b,c,d \in \mathcal{M}$, with the three distance $\rho(\mu(a,b,d),d)$, $\rho(\mu(b,c,d),d)$ and $\rho(\mu(c,a,d),d)$ all bounded, then $\rho(\mu(a,b,c),d)$ is also bounded.)

The argument is essentially the same. This time, we apply Corollary 6.5 to the set A instead of (C2) directly. Suppose that $x,y \in \Pi$ are respectively the π-images of the left and right sides of one of the input identities. As in the previous argument, we see that λx and λy are respectively a bounded distance from the result of applying the same formulae in $\mathcal{M}$, which by assumption, are a bounded distance apart in $\mathcal{M}$. It follows that $\rho(\lambda x, \lambda y)$ is bounded. By choosing the constant t in Corollary 6.5 to be larger than this bound, we see that we must have $x = y$. In other words, this input identity holds exactly in Π, for the π-images of the elements of A. We can assume this is true of all the input identities. Therefore, the derived identity must hold too. Now, again, as in the previous argument, we see that the derived identity holds up to bounded distance in $\mathcal{M}$. This proves the claim.

We can apply the these principles in the following discussion.

Given $a,b \in \mathcal{M}$, we define the *coarse interval* between a and b as $[a,b] = \{\mu(a,b,x) \mid x \in \mathcal{M}\}$. By the observation above, we see that is a bounded Hausdorff distance from $\{x \in M \mid \rho(\mu(a,b,x),x) \le r\}$ for any fixed sufficiently large $r \ge 0$.

Definition. We say that a subset, $C \subseteq \mathcal{M}$, of a coarse median space, $(\mathcal{M},\rho,\mu)$, is *k-(median) quasiconvex* if for all $a,b \in C$ and $x \in \mathcal{M}$, $\rho(\mu(a,b,x),C) \le k$.

From property (C1) we see that any quasiconvex set is quasi-isometrically embedded in $\mathcal{M}$ (or more precisely, some uniform neighbourhood of C is quasi-isometrically embedded with respect to the induced path-metric). Note also quasi-convexity of C is equivalent to asserting that for all $a,b \in C$, the coarse interval $[a,b]$ lies in a uniform neighbourhood of C. Note that Lemma 6.1 implies that if $a,b,c \in \mathcal{M}$ and $c \in [a,b]$, then $\rho(a,b)$ agrees with $\rho(a,c)+\rho(c,b)$ up to linear bounds.

We next recall the following standard notion for any median algebra, M. Suppose that $C \subseteq M$ is (a priori) any subset. We say that a map $f : M \to C$ is a *gate map* if $fx \in [x,c]$ for all $x \in M$ and $c \in C$. Note that if $a,b \in M$ and $c \in [a,b]$ then $fc \in [a,c] \cap [b,c] \in \{c\}$, so $fc = c$. It follows immediately that $f|C$ is the identity (since $c \in [c,c]$), and that C is convex (since $c = fc \in C$). We also claim that f is a homomorphism. For this, it is enough to show that if $c \in [a,b]$, then $fc \in [fa,fb]$. But now the identities $c \in [a,b]$, $fc \in [c,fb]$ and $fb \in [b,fc]$ together imply

$fc \in [a, fb]$. Thus (by the same observation, with a, b, c replaced by fb, a, fc) we get $fc = ffc \in [fa, fb]$ as required. We also note that if a gate map exists for a given C, then it is unique.

We can now define the corresponding notion in a coarse median space, $\mathcal{M}$.

Definition. A map $\phi : \mathcal{M} \to C$ to a subset $C \subseteq \mathcal{M}$ is an *r-coarse gate map* if for all $x \in \mathcal{M}$ and $c \in C$, we have $\rho(x, \mu(x, \phi x, c)) \leq r$.

Lemma 6.6. *If $\phi : \mathcal{M} \to C$ is an r-coarse gate map, then C is k-quasiconvex, ϕ is an l-quasimorphism, and $\rho(c, \phi c) \leq h$ for all $c \in C$, where k, l, h depend only on r and the parameters of $\mathcal{M}$.*

Proof. We follow the same argument as for a median algebra described above, except that now equalities and inclusions are assumed to hold up to bounded distance, depending only on r, and the parameters of $\mathcal{M}$. By the general principles described above, any deduction (based on a finite sequence of identities) in a median algebra holds also in a coarse median algebra, interpreting everything up to bounded distance. □

Suppose now that $((\mathcal{M}_\zeta, \rho_\zeta, \mu_\zeta))_\zeta$ is a $\mathcal{Z}$-sequence of uniformly coarse median spaces (i.e., with parameters independent of ζ). Let $\boldsymbol{t} \in \mathcal{U}\mathbb{R}$ be a positive infinitesimal. We get a limiting space, $(\mathcal{M}^*, \rho^*, \mu^*)$, where $(\mathcal{M}^*, \rho^*)$ is the extended asymptotic cone, and where $(\mathcal{M}^*, \mu^*)$ is a topological median algebra (that is, the map $\mu^* : (\mathcal{M}^*)^3 \to \mathcal{M}^*$ is continuous). If each $\mathcal{M}_\zeta$ has rank at most n (as a coarse median space), then $\mathcal{M}^*$ has rank at most n (as a median algebra). Note that $(\mathcal{M}^*, \mu^*)$ need not be a median metric space, though it satisfies a weaker metric condition described in [Bowditch 2013; 2014a], namely that the maps $x \mapsto \mu(a, b, x)$ are uniformly Lipschitz, for all $a, b \in \mathcal{M}^*$, (see Proposition 9.1 and Lemma 9.2 of [Bowditch 2013]). Note that this is the hypothesis of Proposition 2.4 here.

In those papers, we restricted attention to the asymptotic cone, $\mathcal{M}^\infty$, but that does not affect the above observations.

Lemma 6.7. *Let $\mathcal{M}^*$ be an extended asymptotic cone of a $\mathcal{Z}$-sequence of uniformly coarse median spaces. Suppose that $Q \subseteq \mathcal{M}^*$ is an n-cube. Then we can find a sequence of l_0-quasimorphisms $\phi_\zeta : Q \to \mathcal{M}_\zeta$ such that for all $x \in Q$, $\phi_\zeta x \to x$, where l_0 depends only on n and the uniform parameters of the $\mathcal{M}_\zeta$.*

Proof. To begin, take any sequence of maps, $\psi_\zeta : Q \to \mathcal{M}_\zeta$, with $\psi_\zeta x \to x$ for all $x \in Q$. (Such maps exist directly from the definition of the asymptotic cone.) Since μ^* is, by definition, the limit of the μ_ζ, it follows that ψ_ζ is a h_ζ-quasimorphism, where $t_\zeta h_\zeta \to 0$ (since they must converge to a monomorphism in $\mathcal{M}^*$). Let $\phi_\zeta : Q \to \mathcal{M}_\zeta$ be the l_0-quasimorphism given by Lemma 6.3. For all $x \in Q$, $\rho_\zeta(\phi_\zeta x, \psi_\zeta x) \leq kh_\zeta + h_1$ so $t_\zeta \rho_\zeta(\phi_\zeta x, \psi_\zeta x) \leq kt_\zeta h_\zeta + h_1 t_\zeta \to 0$. Thus $\phi_\zeta x \to x$, as required. □

Note that, if we have a sequence of uniformly quasiconvex sets, $C_\zeta \subseteq \mathcal{M}_\zeta$, we have a limiting bi-Lipschitz embedded closed convex subset $C^* \subseteq \mathcal{M}^*$ in the extended asymptotic cone $\mathcal{M}^*$. If $\phi_\zeta : \mathcal{M}_\zeta \to C_\zeta$ are a sequence of uniform coarse gate maps, the limiting map $\phi^* : \mathcal{M}^* \to C^*$ is a gate map.

As in [Bowditch 2013, Section 12], we say that a median algebra, Π, is *n-colourable* if there is an n-colouring of the walls such that no two walls of the same colour cross. We say that a coarse median space $\mathcal{M}$ is *n-colourable* if in (C2) we can always choose Π to be n-colourable as a median algebra. Clearly this implies that $\mathcal{M}$ has rank at most n. The following was shown in [Bowditch 2014a, Proposition 12.5].

Theorem 6.8. *Suppose that $((\mathcal{M}_\zeta, \rho_\zeta, \mu_\zeta))_\zeta$ is a sequence of n-colourable uniform coarse median spaces, for some fixed n. Then $\mathcal{M}^*$ admits a metric ρ', bi-Lipschitz equivalent to ρ^*, such that $(\mathcal{M}^*, \rho')$ is an (extended) median metric space with median μ^*. Moreover, $\mathcal{M}^*$ is n-colourable as a median algebra.*

In fact, the bi-Lipschitz constant only depends on the parameters of the coarse median spaces.

The construction however is not canonical. Note that the median metric space arising is necessarily proper.

In particular, we see that the asymptotic cone of a sequence of finitely colourable coarse median spaces is bi-Lipschitz equivalent to a proper median metric space, and hence in turn to a CAT(0) space (by Theorem 2.2). In fact, the same holds for a sequence of finite-rank coarse median spaces. This relies on the following variation of Theorem 6.8:

Theorem 6.9. *Suppose that $((\mathcal{M}_\zeta, \rho_\zeta, \mu_\zeta))_\zeta$ is a sequence of coarse median spaces of rank n, for some fixed n. Then $\mathcal{M}^*$ admits an extended metric, ρ', bi-Lipschitz equivalent to ρ^*, such that $(\mathcal{M}^*, \rho')$ is an (extended) median metric space of rank n, with median μ^*.*

Proof. As already observed, it is easily seen from axiom (C1) that ρ^* satisfies the hypotheses of Proposition 2.4, when restricted to any component of M^*. We can therefore apply Proposition 2.4 to each component separately. □

We finish this section by briefly discussing the special case of a Gromov hyperbolic space $(\mathcal{M}, \rho)$. See [Bowditch 2013, Section 3] for elaboration.

Given $a, b, c \in \mathcal{M}$, write

$$\langle a, b{:}c\rangle = \tfrac{1}{2}(\rho(a, c) + \rho(b, c) - \rho(a, b))$$

for the "Gromov product". Up to bounded distance, this is the same as the distance of c to some (or any) geodesic from a to b.

Definition. We say that $m \in \mathcal{M}$ is an *r-centroid* for $a, b, c \in \mathcal{M}$ if $\langle a, b{:}m\rangle \le r$, $\langle b, c{:}m\rangle \le r$ and $\langle c, a{:}m\rangle \le r$.

Provided r is a sufficiently large in relation to the hyperbolicity constant, such an r-centroid will always exist. We will fix such an r and simply refer to m as a *centroid*. In fact, m is well defined up to bounded distance, and we write $\mu(a, b, c) = m$ for some choice of m. With this structure, $(\mathcal{M}, \rho, \mu)$ is a coarse median space of rank at most 1. Indeed, any rank-1 coarse median space arises in this way. (We note that rank-0 is trivially equivalent to having finite diameter.)

If $(\mathcal{M}_\zeta)_\zeta$ is a sequence of uniformly hyperbolic spaces, then $(\mathcal{M}^*, \mu^*)$ is a rank-1 median algebra (variously known in the literature as a "tree algebra", "median pretree", etc.). As already observed in Section 5, $(\mathcal{M}^*, \rho^*)$ is an $\mathbb{R}^*$-tree, and $(\mathcal{M}^\infty, \rho^\infty)$ is an $\mathbb{R}$-tree.

It is shown in [Bowditch 2013] that if $\mathcal{M}$ is a coarse median space of rank at most n, then there is no quasi-isometric embedding of $\mathbb{R}^{n+1}$ into $\mathcal{M}$ (since this would give rise to an injective map of $\mathbb{R}^{n+1}$ into $\mathcal{M}^\infty$, contradicting the fact that $\mathcal{M}^\infty$ has locally compact dimension at most n). In fact, the same argument can be applied to bound the radii of quasi-isometrically embedded balls. To state this more precisely, write B^n_r for the ball of radius r in the euclidean space $\mathbb{R}^n$.

Lemma 6.10. *Let $\mathcal{M}$ be a coarse median space of rank at most n. Given parameters of $\mathcal{M}$ and of quasi-isometry, there is some constant $r \geq 0$, such that there is no quasi-isometric embedding of B^{n+1}_r into $\mathcal{M}$ with these parameters.*

Proof. Suppose that, for each $i \in \mathbb{N}$, the ball, B_i, of radius i admits a uniformly quasi-isometric embedding, $\phi_i : B_i \to \mathcal{M}$. Now pass to the asymptotic cone with indexing set, $\mathcal{Z} = \mathbb{N}$, and with scaling factors $1/i$. We end up with a continuous injective map, $\phi^\infty : B_1 \to \mathcal{M}^\infty$, contradicting the fact that $\mathcal{M}^\infty$ has locally compact dimension at most n.

To see that only the parameters of $\mathcal{M}$ are relevant to the value of r, we should allow $\mathcal{M}$ also to vary among coarse median spaces with these parameters when taking the asymptotic cone. More precisely, suppose we have a sequence, $\phi_i : B_i \to \mathcal{M}_i$, of uniformly quasi-isometric maps, where the $\mathcal{M}_i$ are uniform coarse median spaces. This time, we get a limiting map $\phi^\infty : B_1 \to \mathcal{M}^\infty$, where $\mathcal{M}^\infty$ is the ultralimit of the spaces $(\mathcal{M}_i)_i$ again scaled by $1/i$. This leads to the same contradiction. In other words, there must be a bound on the diameter of a euclidean ball which we can quasi-isometrically embed, for any fixed parameters. □

Note that it follows, for example, that $\mathcal{M}$ admits no quasi-isometrically embedded euclidean half-space of dimension $n + 1$.

Remark. The last paragraph of the proof of Lemma 6.10 is a standard trick to obtain uniform constants and will be used again later. (See the remark after Lemma 14.5.)

7. A general construction of coarse medians

In this section, we give a general criterion for the existence of coarse medians on certain types of spaces associated to a surface. In particular, we will apply this to the marking graph in Section 8, to recover the result of [Behrstock and Minsky 2011]. The argument follows broadly as in that work using [Behrstock, Kleiner, Minsky and Mosher 2012]. In doing this, under our hypotheses, we give a version of the compatibility theorem for medians. To this end, we will list a set of axioms ((A1)–(A10) below) which relate to "projection maps" to spaces indexed by a set, $\mathcal{X}$, namely the collection of "subsurfaces" of a surface Σ. The main results of this section, namely Theorems 7.1 and 7.2, together give a more precise statement of Theorem 1.1.

In Section 8, we will explain how this applies, in particular, to the marking graph, and recover the result of [Behrstock and Minsky 2011].

As mentioned in Section 1, the main purpose of keeping the discussion general is that one can readily check that the hypotheses we give here apply in other situations — notably to Teichmüller space in either the Teichmüller or Weil–Petersson metrics; see [Bowditch 2016a; 2015]. This also applies to most of the discussion in Sections 9–12 here.

We remark that in [Behrstock, Hagen and Sisto 2017], the authors define the notion of a "hierarchically hyperbolic" space, based on a different (though related) set of axioms. These allow for more general indexing sets. However in the case where the indexing set is taken to be $\mathcal{X}$, as in the present paper, one can verify that hierarchically hyperbolic spaces satisfy our axioms, and are hence coarse median. See [Behrstock, Hagen and Sisto 2015, Section 7] for more discussion of this.

Let Σ be a compact orientable surface. Let $\xi(\Sigma)$ be its complexity, that is, $\xi(\Sigma) = 3g + p - 3$, where g is the genus, and p is the number of boundary components. If $\xi(\Sigma) = 0$ then Σ is a three-holed sphere. If $\xi(\Sigma) = 1$ then Σ is a four-holed sphere, or a one-holed torus. We will write $S_{g,p}$ to denote the topological type of surface of genus g and p boundary components.

Definition. By an *essential curve* in Σ, we mean a simple closed curve which homotopically nontrivial and nonperipheral (not homotopic into $\partial\Sigma$). By a *curve* we mean a free homotopy class of essential curves.

Definition. By an *essential subsurface* Σ we mean a compact connected subsurface, $X \subseteq \Sigma$, such that each boundary component of X is either a component of $\partial\Sigma$, or else an essential (and nonperipheral) simple closed curve in $\Sigma \setminus \partial\Sigma$, and such that X is not homeomorphic to a three-holed sphere.

Note that we are allowing Σ itself as a subsurface, as well as nonperipheral annuli.

Definition. A *subsurface* is a free homotopy class of essential subsurfaces.

We refer to an essential surface in the given homotopy class as a *realisation* of the subsurface. Note that there is a natural bijective correspondence between curves and annular subsurfaces.

Given $X, Y \in \mathcal{X}$, we distinguish five mutually exclusive possibilities denoted as follows:

(1) $X = Y$: X and Y are homotopic.

(2) $X \prec Y$: $X \neq Y$, and X can be homotoped into Y but not into ∂Y.

(3) $Y \prec X$: $Y \neq X$, and Y can be homotoped into X but not into ∂X.

(4) $X \wedge Y$: $X \neq Y$, and X, Y can be homotoped to be disjoint.

(5) $X \pitchfork Y$: none of the above.

In (2)–(4) one can find realisations of X, Y in Σ such that $X \subseteq Y$, $Y \subseteq X$, $X \cap Y = \emptyset$, respectively. (Note that $X \wedge Y$ covers the case where X is an annulus homotopic to a boundary component of Y, or vice versa.) We can think of (5) as saying that the surfaces "overlap". We write $X \preceq Y$ to mean $X \prec Y$ or $X = Y$. (Note that this excludes the case where Y is homotopic to an annular boundary component of a nonannular subsurface, X.)

We note that

$$X \wedge Y \Leftrightarrow Y \wedge X, \qquad X \pitchfork Y \Leftrightarrow Y \pitchfork X,$$
$$X \prec Y \prec Z \Rightarrow X \prec Z, \quad X \wedge Y \text{ and } Z \prec Y \Rightarrow X \wedge Z.$$

Given $X \in \mathcal{X}$, write $\mathcal{X}(X) = \{Y \in \mathcal{X} \mid Y \preceq X\}$.

We now introduce the hypotheses of the main result of this section.

We suppose that to each $X \in \mathcal{X}$, we have associated geodesic metric spaces $(\mathcal{M}(X), \rho_X)$ and $(\mathcal{G}(X), \sigma_X)$, as well as a map $\chi_X : \mathcal{M}(X) \to \mathcal{G}(X)$. We will generally abbreviate $\rho = \rho_X$ and $\sigma = \sigma_X$, where there is no confusion. Given $X, Y \in \mathcal{X}$ with $Y \prec X$, we suppose that we have a map $\psi_{YX} : \mathcal{M}(X) \to \mathcal{M}(Y)$. We write $\theta_{YX} = \chi_Y \circ \psi_{YX} : \mathcal{M}(X) \to \mathcal{G}(Y)$. We also assume that if $X, Y \in \mathcal{X}$ with $Y \pitchfork X$, or $Y \prec X$, then we have associated an element $\theta_X Y \in \mathcal{G}(X)$. If α is a curve, we will write $\theta_X \alpha = \theta_X Y$, where $Y = X(\alpha)$ is the annular neighbourhood of α. It will be seen that the hypotheses laid out below only really require these maps to be defined up to bounded distance.

(In Section 8, we will be setting $\mathcal{M}(X) = \mathbb{M}(X)$ and $\mathcal{G}(X) = \mathbb{G}(X)$, to be the intrinsic marking graphs and curve graph respectively, when $X \in \mathcal{X}_N$. The map χ_X is the natural projection, and ψ_{YX} is the usual subsurface projection. If $X \in \mathcal{X}_A$, then $\mathcal{G}(X) = \mathbb{G}(X)$ is the usual graph that measures twisting around the core curve. In this case, we set $\mathbb{M}(X) = \mathbb{G}(X)$ and χ_X to be the identity map.)

We will assume:

(A1) "hyperbolic": There exists $k \ge 0$ such that for all $X \in \mathcal{X}$, $\mathcal{G}(X)$ is k-hyperbolic.

(A2) "χ Lipschitz and cobounded": there exist $k_1, k_2, k_3 \ge 0$ such that for all $X \in \mathcal{X}$ and for all $a, b \in \mathcal{M}(X)$, $\sigma(\chi_X a, \chi_X b) \le \rho(a, b) + k_2$ and $\mathcal{G}(X) = N(\chi_X(\mathcal{M}(X)); k_3)$

(A3) "ψ Lipschitz": there exist $k_1, k_2 \ge 0$ such that for all $X \in \mathcal{X}$ and for all $Y \in \mathcal{X}(X)$ and for all $a, b \in \mathcal{M}(X)$, $\rho(\psi_{YX} a, \psi_{YX} b) \le \rho(a, b) + k_2$.

(A4) "composition": There is some $s_0 \ge 0$ such that if $X, Y, Z \in \mathcal{X}$ with $Z \prec Y \prec X$ and $a \in \mathcal{M}(X)$, then $\rho(\psi_{ZX} a, \psi_{ZY} \circ \psi_{YX} a) \le s_0$.

(A5) "disjoint projection": there exists $s_1 \ge 0$ such that for all $X \in \mathcal{X}$ if $Y, Z \in \mathcal{X}$ with $Y \wedge Z$ or $Y \prec Z$, then $\sigma(\theta_X Y, \theta_X Z) \le s_1$ whenever $\theta_X Y$ and $\theta_X Z$ are defined.

Thus, (A2) and (A3) tell us that our maps are all uniformly coarsely Lipschitz. In view of (A4) we will abbreviate θ_{YX} to θ_Y and ψ_{YX} to ψ_Y, whenever the domain of the map is clear. If $X = Y$, we set $\psi_X = \psi_{XX}$ to be the identity map on $\mathcal{M}(X)$. Note that, with these conventions, we can also write χ_X as θ_X. We will also abbreviate $\sigma_Y(a, b) = \sigma(\theta_Y a, \theta_Y b)$ and $\rho_Y(a, b) = \rho(\psi_Y a, \psi_Y b)$. To simplify the exposition, we will view θ_X as a map from $\mathcal{M}(X) \sqcup \mathcal{X}(X)$ to $\mathcal{G}(Y)$.

Given $a, b \in \mathcal{M}(X)$, write $R_X(a, b) = \max\{\sigma_Y(a, b) \mid Y \in \mathcal{X}(X)\}$. Similarly, if $a \in \mathcal{M}(X)$ and $Z \in \mathcal{X}(X) \setminus \{X\}$, write

$$R_X(a, Z) = \max\{\sigma_Y(a, Z)\}$$

as Y ranges over those elements of $\mathcal{X}(X)$ with either with $Y \prec Z$ or $Y \pitchfork Z$. (In the context of marking graphs, one can view R_X as measuring intersection numbers.)

We assume:

(A6) "finiteness": there exists $r_0 \ge 0$ such that for all $X \in \mathcal{X}$ and for all $a, b \in \mathcal{M}(X)$, the set of $Y \in \mathcal{X}(X)$ with $\sigma_Y(a, b) \ge r_0$ is finite.

(A7) "distance bound": for all $r \ge 0$ there exists $r' \ge 0$ such that for all $X \in \mathcal{X}$ and for all $a, b \in \mathcal{M}(X)$, if $R_X(a, b) \le r$, then $\rho(a, b) \le r'$.

(A8) "bounded image": there exists r_0 such that for all $X \in \mathcal{X}$ and for all $Y \in \mathcal{X}(X)$ and for all $a, b \in \mathcal{M}(X)$, if $\langle \theta_X a, \theta_X b : \theta_X Y \rangle \ge r_0$, then $\sigma_Y(a, b) \le r_0$.

(A9) "overlapping subsurfaces": there exists r_0 such that for all $X \in \mathcal{X}$ and for all $Y, Z \in \mathcal{X}(X)$, if $Y \pitchfork Z$ and $x \in \mathcal{M}(X) \sqcup \mathcal{X}(X)$, then

$$\min\{\sigma_Y(x, Z), \sigma_Z(x, Y)\} \le r_0.$$

(A10) "realisation": there exists r_0 such that for all $X \in \mathcal{X}$, if $\mathcal{Y} \subseteq \mathcal{X}(X)$ with $Y \wedge Z$ for all distinct $Y, Z \in \mathcal{Y}$, and if to each $Y \in \mathcal{Y}$ we have associated some $a_Y \in \mathcal{M}(Y)$, then there is some $a \in \mathcal{M}(Y)$ with $\rho(a_Y, \psi_Y a) \le r_0$ and $R_X(a, Y) \le r_0$ for all $Y \in \mathcal{X}$.

In fact, for (A10) it would be enough to take $\mathcal{Y}$ to consist of an annular subsurface together with any non-$S_{0,3}$ complementary components — we can then keep cutting the surface into smaller and smaller pieces, and the general case follows by an inductive application of (A4) "composition".

Note that, using by (A2) "χ Lipschitz and cobounded" and (A3) "ψ Lipschitz" we have a reverse inequality in (A7) "distance bound", namely that $R_X(a,b)$ is (linearly) bounded above in terms of $\rho(a,b)$.

We note that (A6) "finiteness" and (A7) "distance bound" are both consequences of the following distance formula.

Given $r \geq 0$, we write

$$\mathcal{A}_X(a,b;r) = \{Y \in \mathcal{X}(X) \mid \sigma_Y(a,b) > r\}.$$

Given $a, b \in \mathbb{M}^0(\Sigma)$ and $r \geq 0$, let $D_X(a,b;r) = \sum_{Y \in \mathcal{A}_X(a,b;r)} \sigma_Y(a,b)$.

We suppose:

(B1) "distance formula": there exists $r_0 \geq 0$ such that for all $r \geq r_0$, there exist $k_1 > 0, h_1, k_2, h_2 \geq 0$ such that for all $X \in \mathcal{X}$ and for all $a, b \in \mathcal{M}(X)$,

$$k_1\rho(a,b) - h_1 \leq D_X(a,b;r) \leq k_2\rho(a,b) + h_2.$$

We will sometimes abbreviate this statement to $D_X(a,b;r) \asymp \rho(a,b)$.

Less formally, this says that distances in $\mathcal{M}(X)$ agree to within linear bounds with the sum of all sufficiently large projected distances in $\mathcal{G}(Y)$ as Y ranges over subsurfaces of X. Here "sufficiently large" implies a lower threshold below which we ignore any contributions. The linear bounds will depend on the particular choice of threshold. For this to work, the threshold must be assumed sufficiently large.

In the case of markings, (B1) is the distance formula of [Masur and Minsky 2000], who also stated it for the pants complex. (We remark that for the Teichmüller metric, a similar formula has been proven by Rafi and by Durham, and is used in [Bowditch 2016a]. A more general version, which encompasses these cases is proven in [Behrstock, Hagen and Sisto 2015].) Again for markings, (A8) "bounded image" is a consequence of their bounded geodesic image theorem, (A9) "overlapping subsurfaces" is a consequence of Behrstock's lemma, and (A10) "realisation" is a simple explicit construction. We will elaborate on this in Section 8.

Given $Y \in \mathcal{X}$, we write $\mu_Y : (\mathcal{G}(Y))^3 \to \mathcal{G}(Y)$ for the usual median (or "centroid") operation on the uniformly hyperbolic space $\mathcal{G}(Y)$. (That is, $\mu(a,b,c)$ is a bounded distance from any geodesic connecting any two distinct points of $\{a,b,c\}$.)

We will show:

Theorem 7.1. *Under the hypotheses (A1)–(A10) above, there is some $t_0 \geq 0$ depending only on the parameters of the hypotheses such that if $X \in \mathcal{X}$ and*

$a, b, c \in \mathcal{M}(X)$, *there is some* $m \in \mathcal{M}(X)$ *such that for all* $Y \in \mathcal{X}(X)$ *we have* $\sigma(\theta_Y m, \mu_Y(\theta_Y a, \theta_Y b, \theta_Y c)) \le t_0$.

By (A7) "distance bound", m is well defined up to bounded distance. We set $\mu_X(a, b, c) = m$ for some such m to give us a ternary operation $\mu_X : (\mathcal{M}(X))^3 \to \mathcal{M}(X)$. Using (A4) "composition", we see that if $Y \in \mathcal{X}(X)$, then $\psi_Y : \mathcal{M}(X) \to \mathcal{M}(Y)$ is a uniform quasimorphism, that is,

$$\rho_Y(\mu_Y(\psi_Y a, \psi_Y b, \psi_Y c), \psi_Y \mu_X(a, b, c)) \le h \quad \text{for all } a, b, c \in \mathcal{M}(X),$$

where $h \ge 0$ depends only on the parameters of the hypotheses.

Theorem 7.2. *Under the hypotheses* (A1)–(A10), *there is a ternary operation*, μ_X, *defined on each space* $\mathcal{M}(X)$ *such that* $(\mathcal{M}(X), \rho_X, \mu_X)$ *is a coarse median space*, *and such that the maps* $\theta_{YX} : \mathcal{M}(X) \to \mathcal{G}(Y)$ *for* $Y \preceq X$ *are all median quasimorphisms. The median* μ_X *is unique with this property*, *up to bounded distance. The maps* $\psi_{YX} : \mathcal{M}(X) \to \mathcal{M}(Y)$ *for* $Y \preceq X$ *are also median quasimorphisms. The coarse median space* $(\mathcal{M}(X), \rho_X, \mu_X)$ *is finitely colourable*, *and has rank at most* $\xi(X)$. *Moreover*, *all constants of the conclusion* (*coarse median property*, *and quasimorphism*) *depend only on the constants of the hypotheses* (A1)–(A10).

Proof. Given Theorem 7.1, this follows directly from the results of [Bowditch 2013], in particular, Propositions 10.1 and 10.2 thereof. We just need to check that the respective hypotheses (P1)–(P4) and (P3′) are satisfied.

Here, (P1) is (A7) "distance bound" and (P2) is (A1) "hyperbolic". (P3) is the statement that one can embed at most ξ subsurfaces disjointly in a surface of complexity ξ. Finally, (P4) follows exactly as in Lemma 11.7 there, which only uses properties (A8) "bounded image" and (A9) "overlapping subsurfaces". Moreover, property (P3′) also holds here, for the same reason. ☐

(In some cases, one can improve on the rank bound of Theorem 7.2, as in the case of the Weil–Petersson metric [Bowditch 2015].)

So far, we have made no reference to the action of $\mathrm{Map}(\Sigma)$. In applications, the spaces and maps will be equivariant (up to bounded distance), and it follows that the medians we construct will necessarily also be equivariant up to bounded distance.

We now set about the proof of Theorem 7.1. To simplify the exposition, we will construct the median $\mu = \mu_\Sigma$ on Σ. The same arguments apply working intrinsically in any subsurface $X \in \mathcal{X}_N$.

We begin with some general observations about the treelike (rank-1 median) nature of hyperbolic spaces.

Definition. A *spanning tree* for a finite set A consists of a finite simplicial tree, Δ, and a map $\pi = \pi_\Delta : A \to V(\Delta)$ to the vertex set.

Recall that the vertex set of a finite simplicial tree is a rank-1 median algebra (and every finite rank-1 median algebra has this form). We can assume that every terminal (i.e., degree-1) vertex of Δ lies in πA. We say that Δ is *trivial* if it is a singleton.

Suppose that T is another spanning tree with an embedding of Δ in T. There is a natural retraction, ω, of T onto Δ, and hence of $V(T)$ to $V(\Delta)$. We say that the spanning tree T is an *enlargement* of Δ if $\pi_\Delta = \omega\pi_T$.

Suppose that $\{\Delta_i\}_{i\in\mathcal{J}}$ is a finite collection of spanning trees for A, indexed by some set $\mathcal{J}$. We say that a spanning tree T for A is a *common enlargement* of $\{\Delta_i\}_{i\in\mathcal{J}}$ if we can embed the Δ_i simultaneously in T so that their interiors are disjoint, and such that T is an enlargement of each Δ_i. Note that (after collapsing complementary trees), we may as well suppose that $T = \bigcup_{i\in\mathcal{J}} \Delta_i$. We write $T = T(\{\Delta_i\}_{i\in\mathcal{J}})$. (There may be some ambiguity, in that we may be able to swap to trees each consisting of single edge, and meeting at a vertex not in πA. However, this ambiguity will not matter to us.)

Definition. We say that a collection of spanning trees is *coherent* if it has a common enlargement.

We shall assume henceforth that all our spanning trees are nontrivial.

Lemma 7.3. *Two spanning trees Δ_0 and Δ_1 are coherent if and only if there are vertices, $v_{01} \in V(\Delta_0)$ and $v_{10} \in V(\Delta_1)$ such that $A = \pi_0^{-1}v_{01} \cup \pi_1^{-1}v_{10}$.*

Proof. If $T = T(\Delta_0, \Delta_1)$ is a common spanning tree for A, then T is obtained by taking $\Delta_0 \sqcup \Delta_1$ and identifying a vertex $v_{01} \in V(\Delta_0)$ with $v_{10} \in V(\Delta_1)$, to give a vertex $w \in V(T)$. Note that $\pi : A \to T$ is given by $\pi|(A \setminus \pi_0^{-1}v_{01}) = \pi_0$, $\pi|(A \setminus \pi_1^{-1}v_{10}) = \pi_1$ and $\pi(\pi_0^{-1}v_{01} \cap \pi_1^{-1}v_{10}) = \{w\}$. We can clearly invert the above process. $\square$

Suppose that $\{\Delta_0, \Delta_1, \Delta_2\}$ are coherent. Let $T = T(\Delta_0, \Delta_1, \Delta_2)$. Up to permutation of indices, there are two possibilities:

(1) $\Delta_0, \Delta_1, \Delta_2$ meet at a common vertex $w = V(T)$. In this case, $v_{01} = v_{02}$, $v_{12} = v_{10}$ and $v_{20} = v_{21}$. Note that these vertices all get identified to w in T.

(2) Δ_1 and Δ_2 do not meet in T. In this case, $v_{01} \neq v_{02}$, $v_{12} = v_{10}$ and $v_{20} = v_{21}$.

Note that the conditions on vertices above make sense if we assume only that Δ_0, Δ_1 and Δ_2 are pairwise coherent.

Lemma 7.4. *Let $\{\Delta_0, \Delta_1, \Delta_2\}$ be pairwise coherent. Then it is coherent if an only if at most one of the three equalities $v_{01} = v_{02}$, $v_{12} = v_{10}$ and $v_{20} = v_{21}$ does not hold.*

Proof. We have explained "only if", so we prove "if":

(1) Suppose all the equalities hold. Let $w_0 = v_{01} = v_{02}$, $w_1 = v_{12} = v_{10}$ and $w_2 = v_{20} = v_{21}$. Let T be obtained from $\Delta_0 \sqcup \Delta_1 \sqcup \Delta_2$ by identifying w_0, w_1 and w_2 to a single point $w \in V(T)$. We define $\pi : A \to V(T)$ by $\pi|(A \setminus \pi_i^{-1} w_i) = \pi_i$ for $i = 0, 1, 2$ and setting $\pi(\pi_0^{-1} w_0 \cap \pi_1^{-1} w_1 \cap \pi_2^{-1} w_2) = \{w\}$.

(2) If not, then, without loss of generality, $v_{01} \neq v_{02}$. Let $w_1 = v_{12} = v_{10}$ and $w_2 = v_{20} = v_{21}$. We construct T from $\Delta_0 \sqcup \Delta_1 \sqcup \Delta_2$ by identifying v_{01} with w_1 to give $x_1 \in V(T)$ and v_{02} with w_2 to give $x_2 \in V(T)$. Note that A can be partitioned into five disjoint sets:

$$A_1 = \pi_0^{-1} v_{01} \setminus \pi_1^{-1} w_1, \quad A_{01} = \pi_0^{-1} v_{01} \cap \pi_1^{-1} w_1, \quad A_0 = \pi_1^{-1} w_1 \cap \pi_2^{-1} w_2,$$
$$A_{02} = \pi_0^{-1} v_{02} \cap \pi_2^{-1} w_2, \quad A_2 = \pi_0^{-1} v_{02} \setminus \pi_2^{-1} w_2.$$

We define $\pi : A \to V(T)$ by setting $\pi|A_i = \pi_i$ for $i = 0, 1, 2$ and setting $\pi(A_{01}) = x_1$ and $\pi(A_{02}) = x_2$. □

In fact, three trees are enough: a finite collection of spanning trees for A is coherent if and only if every subset of at most three elements is coherent. This is not hard to verify, but since we won't be needing it, we omit the proof.

We now move on to consider hyperbolic spaces. Recall that

$$\langle x, y{:}z \rangle = \tfrac{1}{2}(\sigma(x, z) + \sigma(y, z) - \sigma(x, y))$$

for the Gromov product.

Lemma 7.5. *Suppose that (G, σ) is k-hyperbolic, $p \in \mathbb{N}$, and $t \geq 0$. Given a set $B \subseteq G$ with $|B| \leq p$, there is a simplicial tree, Δ, and maps $\pi : B \to V(\Delta)$ and $\lambda : V(\Delta) \to G$ such that for all $x, y, z \in V(\Delta)$, if $\langle \lambda x, \lambda y{:}\lambda z \rangle \leq t$, then $z \in [x, y]_{V(\Delta)}$. Moreover, λ is an h-quasimorphism and for all $x \in B$ we have $\sigma(x, \lambda\pi x) \leq h$, where h depends only on k, p and t.*

Proof. This is proven in [Bowditch 2013, Lemma 10.3]. It is a simple consequence of the fact that any finite set of points in a Gromov hyperbolic space can be approximated up to an additive constant by a finite tree (with vertex set B). The additive constant depends only on p and k. For the clause about Gromov products we need to collapse down "short" edges of the tree (hence the dependence of h on s). This can also be phrased in terms of Corollary 6.5 here. (In [Bowditch 2013] we had a stronger condition on the "crossratios" of four points of B, which is easily seen to imply the condition on Gromov products given here.) □

We will apply this to the spaces $\mathcal{G}(X)$ featuring in the hypotheses of Theorem 7.1. By (A1) these are all k-hyperbolic. Recall that we have maps $\theta_X : \mathcal{M}(X) \to \mathcal{G}(X)$.

We fix some $A \subseteq \mathcal{M}(\Sigma)$ with $|A| = p < \infty$. (In our applications here we will have $p \leq 4$, but we can keep the discussion general for the moment.) We will choose universal $t \geq 0$ sufficiently large (depending only on p) as described below. We apply Lemma 7.5 to $B = \theta_X(A) \subseteq \mathcal{G}(X)$ with t as above, to get a tree $\Delta(X)$ and maps $\pi : B \to V(\Delta(X))$ and $\lambda_X = \lambda : V(\Delta(X)) \to \mathcal{G}(X)$. We set $\pi_X = \pi \circ \theta_X : A \to V(\Delta(X))$.

All we require of this until Lemma 7.11, is:

(*) If $a, b, c \in A$ with $\langle \theta_X a, \theta_X b{:}\theta_X c\rangle \leq t$, then $\pi_X c \in [\pi_X a, \pi_X b]_{V(\Delta(X))}$.

In particular, if $\sigma(\theta_X a, \theta_X b) \leq t$, then $\pi_X a = \pi_X b$ (since $\langle \theta_X a, \theta_X a{:}\theta_X b\rangle \leq t$, so $\pi_X b \in [\pi_X a, \pi_X a] = \{\pi_X a\}$). It follows that if $\operatorname{diam}(\theta_X A) \leq t(p)$, then $\Delta(X)$ is trivial (i.e., a singleton).

For future reference (see Lemma 7.11) we also note that λ is an l-quasimorphism, and that for all $a \in A$, $\sigma(\theta_X a, \lambda_X \pi_X a) \leq l$, where $l = h(p)$ depends only on p.

Lemma 7.6. *Let $X, Y \in \mathcal{X}$ with $X \pitchfork Y$, then there are points, $v_{XY} \in V(\Delta(X))$ and $v_{YX} \in V(\Delta(Y))$ such that $A = \pi_X^{-1} v_{XY} \cup \pi_Y^{-1} v_{YX}$.*

Proof. We can assume that neither $V(\Delta(X))$ nor $V(\Delta(Y))$ is trivial. Note that if $a \in A$, with $\sigma(\theta_X a, \theta_X Y) > r_0$, then $\sigma(\theta_Y a, \theta_Y X) \leq r_0$. If this were true for all $a \in A$, we would conclude that $\operatorname{diam}(\theta_Y A) \leq 2r_0 < t(p)$ giving the contradiction that $V(\Delta(Y))$ is trivial. We can thus find $a_{XY} \in A$ with $\sigma(\theta_X a_{XY}, \theta_X Y) \leq r_0$. We set $v_{XY} = \pi_X a_{XY} \in V(\Delta(X))$. We similarly define $v_{YX} = \pi_Y a_{YX} \in V(\Delta(Y))$.

Now suppose that $b \in A \setminus (\pi_X^{-1} v_{XY} \cup \pi_Y^{-1} v_{YX})$. Then $\pi_X b \neq \pi_X a_{XY}$, and so $\sigma(\theta_X b, \theta_X a_{XY}) \geq t(p)$. Thus, $\sigma(\theta_X b, \theta_X Y) \geq t(p) - r_0 > r_0$. Similarly, $\sigma(\theta_Y b, \theta_Y X) > r_0$. This contradicts property (A9) "overlapping subsurfaces", proving that no such b exists. □

Note that, by Lemma 7.3, we can naturally combine $\Delta(X)$ and $\Delta(Y)$ into a larger tree by identifying the vertices v_{XY} and v_{YX}. In other words, $\{\Delta(X), \Delta(Y)\}$ is coherent. We write $\Delta(X, Y)$ for the common enlargement.

Note that by construction if $\Delta(X)$ and $\Delta(Y)$ are nontrivial, then $\sigma(\theta_X a_{XY}, \theta_X Y) \leq r_0$, where $a_{XY} \in A$ is as in the proof of Lemma 7.6. By the same argument, if $Z \in \mathcal{X}$ with $\Delta(Z)$ nontrivial, we have $\sigma(\theta_X a_{XZ}, \theta_X Z) \leq r_0$, for some $a_{XZ} \in A$. If $\sigma(\theta_X Y, \theta_X Z) < t - 2r_0$, then $\sigma(\theta_X a_{XY}, \theta_X a_{XZ}) < s(p)$, so $v_{XY} = \pi_X a_{XY} = \pi_X a_{XZ} = v_{XZ}$. For future reference (Lemma 7.11) we also note that

$$\sigma(\theta_X a_{XY}, \lambda_X v_{XY}) = \sigma(\theta_X a_{XY}, \lambda_X \pi_X a_{XY}) \leq l,$$

so $\sigma(\theta_X Y, \lambda_X v_{XY}) \leq r_0 + l$).

We write $\mathcal{X}_0$ for the set of $X \in \mathcal{X}$ such that $\Delta(X)$ is nontrivial. It follows from property (A6) "finiteness", that $\mathcal{X}_0$ is finite.

Note that if $X, Y \in \mathcal{X}_0$ and $X \pitchfork Y$, then $\{\Delta(X), \Delta(Y)\}$ is coherent. This is an immediate consequence of Lemmas 7.4 and 7.6. Note that this determines vertices $v_{XY} \in \Delta(X)$ and $v_{YX} \in \Delta(Y)$ which get identified in $\Delta(X, Y)$.

Lemma 7.7. *Suppose that $X, Y, Z \in \mathcal{X}_0$ and that $X \pitchfork Y$ and $X \pitchfork Z$ and $v_{XY} \neq v_{XZ}$. Then $Y \pitchfork Z$.*

Proof. If not, then (since there must be boundary curves of Y and Z which are disjoint) and by (A5) "disjoint projection" we must have $\sigma(\theta_X Y, \theta_X Z) \leq r$, for some constant r depending only on that of (A5), which in turn depends only (or at most) on $\xi(\Sigma)$. Provided we have chosen $t > l + 2r_0$, this implies that $v_{XY} = v_{XZ}$. □

Lemma 7.8. *Suppose that $X, Y, Z \in \mathcal{X}_0$ and that $X \pitchfork Y$, $X \pitchfork Z$ and $Y \pitchfork Z$. Then $\{\Delta(X), \Delta(Y), \Delta(Z)\}$ is coherent.*

Proof. By Lemma 7.7, it's enough to show that at least two of $v_{XY} = v_{XZ}$, $v_{YZ} = v_{YX}$, $v_{ZX} = v_{ZY}$ must hold.

By property (A9) "overlapping subsurfaces",

$$\min\{\sigma(\theta_X Y, \theta_X Z), \sigma(\theta_Y X, \theta_Y Z)\} \leq r_0.$$

Therefore, if $t \geq 3r_0$, we see that either $v_{XY} = v_{XZ}$ or $v_{YZ} = v_{YX}$. Similarly, we have ($v_{YZ} = v_{YX}$ or $v_{ZX} = v_{ZY}$) and ($v_{ZX} = v_{ZY}$ or $v_{XY} = v_{XZ}$), and so the statement follows. □

We can now start on the proof of Theorem 7.1

Suppose $a, b, c \in \mathcal{M}(\Sigma)$. We want to find a median for a, b, c in $\mathcal{M}(\Sigma)$. First choose any $d \in \mathcal{M}(\Sigma)$ with $\sigma_\Sigma(\theta_\Sigma d, \mu_\Sigma(\theta_\Sigma a, \theta_\Sigma b, \theta_\Sigma c))$ bounded in $\mathcal{G}(\Sigma)$. (Such a d exists, since $\chi_\Sigma(\mathcal{M}(\Sigma))$ is cobounded in $\mathcal{G}(\Sigma)$ by (A2) "χ Lipschitz an cobounded".)

Now set $A = \{a, b, c, d\}$, and let $\pi_X : A \to \Delta(X)$ be as described in Lemma 7.5. Let $h = h(4)$. Write $d_X = \pi_X d$ and $e_X = \mu_X(\pi_X a, \pi_X b, \pi_X c)$. Recall that $\mathcal{X}_0$ is the (finite) set of $X \in \mathcal{X}$ such that $\Delta(X)$ is nontrivial. Let $\mathcal{X}_1 = \{X \in \mathcal{X}_0 \mid e_X \neq d_X\}$. By the choice of d, we see that $\Sigma \notin \mathcal{X}_1$.

Suppose that $X, Y \in \mathcal{X}_1$ with $X \pitchfork Y$. Recall that $T = \Delta(X, Y)$ is obtained by identifying $v_{XY} \in \Delta(X)$ with $v_{YX} \in \Delta(Y)$, to give $w \in T$. Note that $\pi_T d$ and $\mu_T(\pi_T a, \pi_T b, \pi_T c)$ must be distinct from w, and must lie in different subtrees $\Delta(X)$ and $\Delta(Y)$. It follows that exactly one of the following must hold:

(1) $d_Y = v_{YX}$ and $e_X = v_{XY}$, or

(2) $d_X = v_{XY}$ and $e_Y = v_{YX}$.

We write these cases respectively as $X \ll Y$ and $Y \ll X$ (which we take to imply that $X \pitchfork Y$).

(Intuitively, we think of these relations as follows. We imagine any finite set of elements of $\mathcal{X}$ embedded disjointly as "horizontal" surfaces in $\Sigma \times \mathbb{R}$; that is, $X \in \mathcal{X}$ is identified with $X \times \{x\}$ for some $x \in \mathbb{R}$. The relations $=$, $\prec$, $\wedge$ and $\pitchfork$ have their usual meaning on projecting to Σ, and $X \ll Y$ means that $X \pitchfork Y$ and X is "to the left" of Y in the sense that it has smaller $\mathbb{R}$-coordinate. The relations are well defined up to isotopy, and satisfy the same properties as those laid out here. This picture ties in with the Minsky model for hyperbolic 3-manifolds homeomorphic to $\Sigma \times \mathbb{R}$.)

Lemma 7.9. *If $X, Y, Z \in \mathcal{X}_1$ and $X \ll Y$ and $Y \ll Z$, then $X \ll Z$.*

Proof. Since $X \ll Y$, $v_{XY} = d_Y$. Since $Y \ll Z$, $v_{XY} = d_Y$. Since $Y \in \mathcal{X}_1$, $d_Y \neq e_Y$, so $v_{YX} \neq v_{YZ}$. By Lemmas 7.7 and 7.8, $X \pitchfork Z$, and $\{\Delta(X), \Delta(Y), \Delta(Z)\}$ is coherent. In particular, $e_X = v_{XY} = v_{XZ}$ and $d_Z = v_{ZY} = v_{ZX}$ so $X \ll Z$. □

Recall that $X \prec Y$ implies that $X \neq Y$ and X is homotopic into Y. We therefore have two strict partial orders $\ll$ and $\prec$ on $\mathcal{X}_1$. Moreover, by hypothesis, $X \ll Y$ is incompatible with any of $X \prec Y$, $Y \prec X$, or $X \wedge Y$.

Lemma 7.10. *Given $X, Y, Z \in \mathcal{X}_1$ with $X \ll Y$ and $Y \prec Z$, then either $X \ll Z$ or $X \prec Z$.*

Proof. Recall that $X \pitchfork Z$ implies $X \ll Z$ or $Z \ll X$. Thus, if the conclusion of the lemma fails, the only alternatives would be $Z = X$, $Z \prec X$, $Z \ll X$ or $Z \wedge X$. Now $Z = X$ or $Z \prec X$ both give $Y \prec X$ contradicting $X \ll Y$; $Z \ll X$ gives $Z \ll Y$ contradicting $Y \prec Z$, and finally, $Z \wedge X$ gives $Y \wedge X$, contradicting $X \ll Y$. □

Now write $X < Y$ to mean that either $X \ll Y$ or $X \prec Y$. This relation is antisymmetric on $\mathcal{X}_1$. It is not in general transitive, but in view of Lemma 7.10, any relation of the form $X < Y < Z < W$ can be reduced to $X < V < W$ for $V \in \{Y, Z\}$. In particular, there are no cycles. It follows that $\mathcal{X}_1$ contains an element U which is maximal with respect to this relation. In other words, if $X \in \mathcal{X}_1$, then we have neither $U \ll X$ nor $U \prec X$. Note that $\Sigma \notin \mathcal{X}_1$, so $U \neq \Sigma$.

From this, we can deduce:

Lemma 7.11. *There is some universal $u_0 > 0$, such that if $a, b, c \in \mathcal{M}(\Sigma)$, there is some curve α such that if $X \in \mathcal{X}$, with $\alpha \pitchfork X$ or $\alpha \prec X$, then*

$$\sigma(\theta_X \alpha, \mu_X(\theta_X a, \theta_X b, \theta_X c)) \leq u_0.$$

Proof. Let $U \in \mathcal{X}_1$ be maximal with respect to $<$, as above. Let α be a component of the relative boundary of U in Σ. Suppose that $X \in \mathcal{X}$ with $\alpha \prec X$ or $\alpha \pitchfork X$. Then either $U \prec X$ or $U \pitchfork X$. According to the conventions described in Section 6, we use the notation $\sim$ to mean "up to bounded distance". In all cases, $\theta_X \alpha$ is defined and $\theta_X \alpha \sim \theta_X U$, by (A5) "disjoint projection". Let $\lambda_X : V(\Delta_X) \to \mathcal{G}(X)$ be the quasimorphism described above (as given by Lemma 7.5). Now,

$$\begin{aligned}
\lambda_X e_X &= \lambda_X \mu_{V(\Delta(X))}(\pi_X a, \pi_X b, \pi_X c) \\
&\sim \mu_X(\lambda_X \pi_X a, \lambda_X \pi_X b, \lambda_X \pi_X c) \\
&\sim \mu_X(\theta_X a, \theta_X b, \theta_X c).
\end{aligned}$$

We therefore want to show that $\theta_X U \sim \lambda_X e_X$. Note that $\lambda_X d_X = \lambda_X \pi_X d \sim \theta_X d$.

Suppose first that $U \prec X$. Thus $X \notin \mathcal{X}_1$, so $d_X = e_X$. Now $\lambda_X e_X = \lambda_X d_X$, is a centroid for $\theta_X a, \theta_X b, \theta_X c$ in $\mathcal{G}(X)$, and so $\theta_X d$ is also a centroid. Therefore, if $\theta_X U$ were far enough away from $\theta_X d$, (depending only on the hyperbolicity constant), then we can assume that the Gromov products $\langle \theta_X a, \theta_X b{:}\theta_X U\rangle$ and $\langle \theta_X a, \theta_X d{:}\theta_X U\rangle$ are both greater than r_0 (after permuting a, b, c as necessary). By property (A8) "bounded image", this implies that $\sigma_U(\theta_U a, \theta_U b) \le r_0$ and $\sigma_U(\theta_U a, \theta_U d) \le r_0$. It then follows that $\pi_U a = \pi_U b = \pi_U d \in V(\Delta(U))$, so $e_U = \mu_{V(\Delta(U))}(\pi_U a, \pi_U b, \pi_U c) = \pi_U d = d_U$, contradicting the fact that $U \in \mathcal{X}_1$. We have shown that if $U \prec X$, then $\theta_X U \sim \lambda_X e_X$ as required.

Suppose now that $U \pitchfork X$. In this case, by Lemma 7.6, the trees $\Delta(X)$ and $\Delta(U)$ are coherent. Moreover, since $\Delta(U)$ is nontrivial, we have $\theta_X U \sim \lambda_X v_{XU}$. If $X \in \mathcal{X}_1$, then $X \ll U$, so $e_X = v_{XU}$, thus $\theta_X U \sim \lambda_X v_{XU} = \lambda_X e_X$ as required. So we can suppose that $X \notin \mathcal{X}_1$ — in other words, $d_X = e_X$. If $X \notin \mathcal{X}_0$, then $\Delta(X)$ is trivial, so $e_X = d_X = v_{XU}$, and we are done, as above. If $X \in \mathcal{X}_0$, then again $d_X = v_{XU}$, otherwise we would get $e_U = d_U$ contradicting $U \in \mathcal{X}_1$.

In all cases, we have shown that $\theta_X U \sim \lambda_X e_X$, as required. □

We can now prove the main result of this section:

Proof of Theorem 7.1. Uniqueness up to bounded distance is an immediate consequence of property (A7) "distance bound" here, so we prove existence.

Let $a, b, c \in \mathcal{M}(\Sigma)$. Let α be a curve as given by Lemma 7.11. Given $X \in \mathcal{X}$, write $\delta_X = \mu_X(\theta_X a, \theta_X b, \theta_X c) \in \mathcal{G}(X)$. So $\theta_X \alpha \sim \delta_X$ for all X with $\alpha \pitchfork X$ or $\alpha \prec X$. We consider only the case when α separates Σ. The nonseparating case is essentially the same.

Let $\Sigma = Y \cup Z$, where $Y \cap Z = \alpha$. Suppose first that neither Y nor Z is a $S_{0,3}$, so that $Y, Z \in \mathcal{X}$. By induction on the complexity of Σ, we can assume that Theorem 7.1 holds intrinsically to Y and Z. Thus, we can find $m_Y \in \mathcal{M}(Y)$, such that if $X = Y$ or $X \prec Y$, then $\sigma(\theta_X m_Y, \delta_X)$ is bounded. We have a similar element, $m_Z \in \mathcal{M}(Z)$. Let $\Omega \in \mathcal{X}_A$ be the annulus with core curve α. We apply property (A10) "realisation" with $\mathcal{Y} = \{X, Y, \Omega\}$ to give $m \in \mathcal{M}(\Sigma)$ such that $\rho(\psi_Y m, m_Y)$, $\rho(\psi_Z m, m_Z)$ and $\rho_\Omega(\psi_\Omega m, \delta_\Omega)$ are bounded. By (A4) "composition" and the construction of m_Y and m_Z, we have $\theta_X m \sim \delta_X$ for all $X \preceq Y$ all $X \preceq Z$.

Suppose that $X \in \mathcal{X}$. If $X \preceq Y$, $X \preceq Z$ or $X = \Omega$, then $\sigma(\theta_X m, \delta_X)$ is bounded by construction. If not, then either $\alpha \prec X$ or $\alpha \pitchfork X$. But then, by the choice of α, $\sigma(\theta_X \alpha, \delta_X)$ is bounded as already observed. But $\sigma(\theta_X m, \theta_X \alpha) \le R_\Sigma(m, a)$ is bounded by (A10) "realisation", so we are done in this case.

If either Y or Z is an $S_{0,3}$, we just omit that subsurface from $\mathcal{Y}$, and proceed in the same way.

If α does not separate, we set $\mathcal{Y}$ to consist of $X(\alpha)$ together with complement of α and proceed similarly. □

8. The marking complex

In this section, we apply the results of Theorem 7.1 to the marking complex of Σ, to recover the result of [Behrstock and Minsky 2011], stated as Theorem 8.2 here. We first describe the curve graph associated to a compact surface, Σ.

For $\xi(\Sigma) \geq 1$, let $\mathbb{G} = \mathbb{G}(\Sigma)$ be the curve graph of Σ. Its vertex set, $\mathbb{G}^0$, is the set of free homotopy classes of essential nonperipheral simple closed curves in Σ. As before, we refer to elements of $\mathbb{G}^0$ simply as *curves*. Two curves, $\alpha, \beta \in \mathbb{G}^0$ are adjacent if $\iota(\alpha, \beta)$ is equal to 2 if Σ is an $S_{0,4}$; equal to 1 if Σ is an $S_{1,1}$; or equal to 0 if $\xi(\Sigma) \geq 2$. Here $\iota(\alpha, \beta)$ denotes the geometric intersection number.

In all cases, $\mathbb{G}(\Sigma)$ is connected. A key result in the subject is:

Theorem 8.1. *There is a universal constant k such that for any compact surface Σ, $\mathbb{G}(\Sigma)$ is k-hyperbolic.*

The existence of such a k, depending on $\xi(\Sigma)$, was proven by Masur and Minsky [1999]. The fact that it is uniform (independent of $\xi(\Sigma)$) was proven independently in [Aougab 2013; Bowditch 2014b; Clay, Rafi and Schleimer 2014; Hensel, Przytycki and Webb 2015]. (The uniformity is not essential to the main results of this paper: we will only be dealing with finitely many topological types at any given time, namely subsurfaces of a given surface Σ. One can therefore simply assert dependence of constants on $\xi(\Sigma)$ at the relevant points.)

Given nonempty $a, b \subseteq \mathbb{G}^0$, let $\iota(a, b) = \max\{\iota(\alpha, \beta) \mid \alpha \in a,\ \beta \in b\}$. We write $\iota(a) = \iota(a, a)$.

Definition. If $\iota(a) = 0$, we refer to a as a *multicurve*.

Intuitively, we think of a multicurve in terms of its realisation as 1-manifold in M.

Definition. We say that $a \subseteq \mathbb{G}^0$ *fills* Σ if $i(a, \gamma) \neq 0$ for all $\gamma \in \mathbb{G}^0$.

If we realise a minimally, then this is the same as saying that all complementary components of $\bigcup a$ are disc or peripheral annuli.

Given $p, q \in \mathbb{N}$, define a graph $\mathbb{M} = \mathbb{M}(\Sigma, p, q)$ by taking the vertex set, $\mathbb{M}^0$ to be the set of $a \subseteq \mathbb{G}^0$ such that a fills Σ and $\iota(a) \leq p$, and by deeming $a, b \in \mathbb{M}^0$ to be adjacent if $\iota(a, b) \leq q$. This graph is always locally finite. Provided p is large enough and q is large enough in relation to p (independently of Σ) it will always be nonempty and connected. For definiteness, we can set $\mathbb{M}(\Sigma) = \mathbb{M}(\Sigma, 2, 4)$, though the actual choice will not matter. (The inclusion of $\mathbb{M}(\Sigma, 2, 4)$ into any larger $\mathbb{M}(\Sigma, p, q)$ is a quasi-isometry.)

Definition. We refer to $\mathbb{M}(\Sigma)$ as the *marking graph* of Σ.

(This is a slight variation on the marking complex of [Masur and Minsky 2000].)

Note that the mapping class group $\mathrm{Map}(\Sigma)$ acts on $\mathbb{G}(\Sigma)$ and on $\mathbb{M}(\Sigma)$ with finite quotient. In particular, we see that $\mathrm{Map}(\Sigma)$ is quasi-isometric to $\mathbb{M}(\Sigma)$. Note also that bounding distance in the marking complex is equivalent to bounding intersection numbers between markings.

Recall that $\mathcal{X} = \mathcal{X}_A \sqcup \mathcal{X}_N$ is the set of (non-$S_{0,3}$) subsurfaces of Σ, partitioned into annular and nonannular subsurfaces.

If $X \in \mathcal{X}_N$, we can define $\mathbb{G}(X)$ and $\mathbb{M}(X)$ intrinsically to X as above. If $X \in \mathcal{X}_A$, one needs to define $\mathbb{G}(X)$ as an arc complex in the annular cover of Σ corresponding to X; see [Masur and Minsky 2000, Section 2.4]. This is quasi-isometric to the real line. In this case, we set $\mathbb{M}(X) = \mathbb{G}(X)$. (One could give a unified description in terms of covers of Σ corresponding to subsurfaces, though we will omit discussion of that here.) We will write $\mathbb{G}(\gamma) = \mathbb{G}(X)$ and $\mathbb{M}(\gamma) = \mathbb{M}(X)$, when $\gamma \in \mathbb{G}^0$, where $X = X(\gamma)$ is the annular neighbourhood of γ.

We will write $\sigma = \sigma_X$ and $\rho = \rho_X$ respectively for the combinatorial metrics on $\mathbb{G}(X)$ and $\mathbb{M}(X)$.

Given $X \in \mathcal{X}$ we have a map $\chi_X : \mathbb{M}(X) \to \mathbb{G}(X)$. If $X \in \mathcal{X}_A$, this is the identity. If $X \in \mathcal{X}_N$, it just chooses some curve from the marking. Up to bounded distance, the map χ_X is determined by the fact that $\iota(a, \chi_X a)$ is bounded for all $a \in \mathbb{M}(X)$.

Given $X, Y \in \mathcal{X}$ with $Y \preceq X$, we have a subsurface projection $\psi_{YX} : \mathbb{M}(X) \to \mathbb{M}(Y)$. This is the same construction as in [Masur and Minsky 2000]. We realise a and Y in minimal general position (so that $a \cap Y$ has a minimal number of components). Now $a \cap Y$ consists of a collection of arcs and curves. We say two arcs are "parallel" if they are homotopic, sliding the endpoints in the boundary components of Y. For each parallel class of arcs we get a disjoint curve (namely the boundary component of a regular neighbourhood of the arc union the boundary components it meets). The collection of such curves, together with the curves of a already lying in Y, give us a collection of curves of Y of bounded self-intersection, and hence give rise to a marking of Y. We write this as $\psi_{YX} a$. Up to bounded distance, the map ψ_{YX} is determined by the fact that the intersection of $\psi_{YX} a$ with every component of $a \cap X$ is bounded.

We set $\theta_Y = \chi_Y \circ \psi_{YX} : \mathbb{M}(X) \to \mathbb{G}(Y)$.

One can also define subsurface projection for curves. Suppose $\gamma \in \mathbb{G}^0(\Sigma)$ and $X \in \mathcal{X}$ with $\gamma \pitchfork X$ or $\gamma \prec X$, then we can define $\theta_X(\gamma) \in \mathbb{G}(X)$. This is consistent with that already defined, in that if $\gamma \in a \in \mathbb{M}^0(X)$, then $\theta_X(\gamma) \sim \theta_X(a)$. In particular, $\theta_X \circ \chi_X(a) \sim \theta_X(a)$ when this is defined. Similarly, if $X, Y \in \mathcal{X}$ with $Y \pitchfork X$ or $Y \prec X$ we can define $\theta_X(Y) \in \mathbb{G}(X)$. This can be defined by setting $\theta_X(Y) = \theta_X(\gamma)$ for some boundary curve, γ, of Y.

We can now deduce the following result [Behrstock and Minsky 2011]:

Theorem 8.2. *There is a constant t_0, depending only on $\xi(\Sigma)$, such that if $a, b, c \in \mathbb{M}(\Sigma)$, then there is some $m \in \mathbb{M}(\Sigma)$ such that for all $X \in \mathcal{X}(\Sigma)$,*

$$\sigma(\theta_X m, \mu_X(\theta_X a, \theta_X b, \theta_X c)) \le t_0.$$

Moreover, if $m' \in \mathbb{M}(\Sigma)$ *is another such element, then* $\rho(m, m') \le t_1$, *where* t_1 *is a constant depending only on* $\xi(\Sigma)$.

We can therefore define a median map $\mu : (\mathbb{M}(\Sigma))^3 \to \mathbb{M}(\Sigma)$ by $\mu(a, b, c) = m$. Of course, it is enough to define $\mu(a, b, c)$ for a, b, c in the vertex set, $\mathbb{M}^0(\Sigma)$, of $\mathbb{M}(\Sigma)$.

To prove Theorem 8.2, we set $\mathcal{M}(X) = \mathbb{M}(X)$ and $\mathcal{G}(X) = \mathbb{G}(X)$. We verify (A1)–(A10) of Section 7 for these spaces and the maps χ_X and ψ_{YX} defined above. In fact, for (A6) "finiteness" and (A7) "distance bound" we will verify (B1) "distance formula".

Note that (A1) "hyperbolicity" is an immediate consequence of Theorem 8.1 above. Properties (A2) "χ Lipschitz and cobounded", (A3) "ψ Lipschitz" and (A4) "composition" are elementary properties of subsurface projection, and (A5) "disjoint projection" holds with $s_1 = 1$ in this case.

Property (B1) "distance formula" is an immediate consequence of the following due to Masur and Minsky [2000] (applied intrinsically to subsurfaces).

Theorem 8.3 [Masur and Minsky 2000]. *There is some* $r_0 \ge 0$ *depending only on* $\xi(\Sigma)$ *such that for all* $r \ge r_0$, *there are constants*, $k_1 > 0$, $h_1, k_2, h_2 \ge 0$ *depending only on* r *and* $\xi(\Sigma)$ *such that if* $a, b \in \mathbb{M}^0(\Sigma)$, *then* $k_1\rho(a, b) - h_1 \le D_\Sigma(a, b; r) \le k_2\rho(a, b) + h_2$.

This implies (A6) and (A7). Property (A8) "bounded image" is an immediate consequence of the bounded geodesic image theorem; [Masur and Minsky 2000, Theorem 3.1]. (Note that the Gromov product $\langle \alpha, \beta{:}\gamma \rangle_X$ is, up to an additive constant, the same as the distance from γ to any geodesic from α to β.) A simpler proof of the bounded geodesic image theorem (with uniform constants, independent of $\xi(\Sigma)$) is given in [Webb 2015].

Property (A9) "overlapping subsurfaces" is an immediate consequence of Behrstock's lemma:

Lemma 8.4. *There is some universal* r_0 *such that if* $X, Y \in \mathcal{X}$ *and* $\gamma \in \mathbb{G}^0(\Sigma)$ *with* $X \pitchfork Y$, $\gamma \pitchfork X$ *and* $\gamma \pitchfork Y$, *then* $\min\{\sigma(\theta_X(\gamma), \theta_X(Y)), \sigma(\theta_Y(\gamma), \theta_Y(X))\} \le r_0$.

This is Theorem 4.3 of [Behrstock 2006] (where r_0 may depend on $\xi(\Sigma)$). A simpler proof, which gives explicit universal constants can be found in [Mangahas 2010].

Property (A10) "realisation" is a simple explicit construction. We can assume that $X = \Sigma$. Let τ be the multicurve consisting of the union of the $\partial_\Sigma Y$, as Y ranges over $\mathcal{Y}$. Each marking m_Y for $Y \in \mathcal{Y} \cap \mathcal{X}_N$ gives us a marking on some component of $\Sigma \setminus \tau$. We now take the union of τ with the union of all these markings. We add in curves transverse to each of the elements of τ to give us a set of curves which fill Σ with bounded self-intersection. We can arrange (after applying suitable Dehn twists about the elements of τ) that the marking has the correct projection to the elements of $\mathcal{Y} \cap \mathcal{X}_A$. Note that this construction automatically gives us a marking, m, with $\iota(m, \tau)$ bounded. By construction, $\psi_Y \sim m_Y$ for all $Y \in \mathcal{Y}$. Also if $Z \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with $Y \prec Z$ or $Y \pitchfork Z$, then $\theta_Z m \sim \theta_Z \tau \sim \theta_Z Y$, so $\sigma_Z(m, Y) \sim 0$, and it follows that $R_\Sigma(m, Y) \sim 0$ as required for (A10).

Finally note that bounding the distance, $\rho(a, b)$ between two markings, $a, b \in \mathbb{M}^0(\Sigma)$ is equivalent to bounding their intersection number, $\iota(a, b)$, which in turn, is equivalent to bounding the quantity, $R_\Sigma(a, b)$ featuring in (A7) "distance bounds". One can find explicit estimates in the references cited, though we will not need them here.

9. Multicurves

Again, in this section, Σ will be a compact surface with $\xi(\Sigma) \geq 1$. We will again assume the hypotheses of Section 7, as we recall below.

Let $\tau \subseteq \mathbb{G}^0(\Sigma)$ be a multicurve in Σ. As usual, we will often identify τ with its realisation as a 1-manifold in Σ. Let $\mathcal{X}_A(\tau) = \{X(\gamma) \in \mathcal{X}_A \mid \gamma \in \tau\}$ be the set of annular surfaces corresponding to the components of τ. Let $\mathcal{X}_N(\tau) \subseteq \mathcal{X}_N$ be the set of components of $\Sigma \setminus \tau$ which are not the $S_{0,3}$. We write $\mathcal{X}(\tau) = \mathcal{X}_A(\tau) \sqcup \mathcal{X}_N(\tau)$.

Given $Y \in \mathcal{X}$, we write $\tau \pitchfork Y$ to mean that $\gamma \pitchfork Y$ or $\gamma \prec Y$ for some $\gamma \in \tau$. Let $\mathcal{X}_T(\tau) = \{Y \in \mathcal{X} \mid Y \pitchfork \tau\}$. Let $\mathcal{X}_I(\tau) = \mathcal{X} \setminus \mathcal{X}_T(\tau)$. It is easily seen that $Y \in \mathcal{X}_I(\tau)$ if and only if $Y \preceq X$ for some $X \in \mathcal{X}(\tau)$. In other words, Y can be homotoped into some component of $\Sigma \setminus \tau$. (This includes the possibility that Y is homotopic to a component of τ.)

Now suppose we have spaces $\mathcal{G}(X)$, $\mathcal{M}(X)$ and maps ψ_{YX}, χ_X, θ_X, etc., satisfying the hypotheses (A1)–(A10) of Section 7. We refer to the constants featuring in these axioms as the *parameters* of $\mathcal{M}(\Sigma)$.

Given $X \in \mathcal{X}_T(\tau)$, we set $\theta_X(\tau) = \theta_X(\gamma)$ for some $\gamma \in \tau$. By (A5) "disjoint projection", we have $\sigma(\theta_X \tau, \theta_X Y) \leq s_1$ for all $Y \in \mathcal{X}_I(\tau)$. In particular, $\theta_X(\tau)$ is well defined up to bounded distance. As usual, we will abbreviate $\sigma_X(\tau, a) = \sigma(\theta_X \tau, \theta_X a)$ for $a \in \mathcal{M}(\Sigma)$ etc.

Given $a \in \mathcal{M}(\Sigma)$, write

$$R(a, \tau) = \max\{\sigma_X(a, \tau) \mid X \in \mathcal{X}_T(\tau)\}.$$

Thus $R(a, \tau) = \max\{R(a, \gamma) \mid \gamma \in \tau\}$ (cf., the definition of R_Σ in Section 7). (In the case of markings, one can think of this as measuring intersection numbers; see Lemma 9.7.) Given $r \geq 0$, let

$$T(\tau; r) = \{a \in \mathcal{M}(\Sigma) \mid R(a, \tau) \leq r\}.$$

Note that if $a \in T(\tau; r)$ and $Y \in \mathcal{X}_I(\tau)$, then $\sigma(a, Y) \le r + s_1$ (by (A10) "realisation"). Also, if for each $X \in \mathcal{X}(\tau)$, we have associated some $a_X \in \mathcal{M}(X)$; then by (A10), there is some $a \in T(\tau; r_0')$ with $\rho(a_X, \psi_X a) \le r_0$ for all $X \in \mathcal{X}_T(\tau)$, where $r_0' = r_0 + s_1$. Note that a is well defined up to bounded distance. In fact:

Lemma 9.1. *If $a, b \in T(\tau; r)$ with $\rho_X(a, b) \le r'$ for all $X \in \mathcal{X}(\tau)$, then $\rho(a, b)$ is bounded above in terms of r and r'.*

Proof. Suppose that $Y \in \mathcal{X}$. If $Y \in \mathcal{X}_I(\tau)$, then $\sigma_Y(a, b)$ is bounded using (A5) "disjoint projection". If $Y \in \mathcal{X}_T(\tau)$, the $\sigma_Y(a, \tau)$ and $\sigma_Y(b, \tau)$ are both bounded above by hypothesis, so $\sigma_Y(a, b)$ is again bounded. The statement now follows by (A7) "distance bound". □

We will abbreviate $T(\tau) = T(\tau; r_0')$.

Let $\mathcal{T}(\tau) = \prod_{X \in \mathcal{X}(\tau)} \mathcal{M}(X)$. We give $\mathcal{T}(\tau)$ the l^1 metric (though any quasi-isometrically equivalent metric would serve for our purposes). Note that $\mathcal{T}(\tau)$ is a coarse median space, with the median defined coordinatewise. We can combine the maps $\psi_X : \mathcal{M}(\Sigma) \to \mathcal{M}(X)$ for $X \in \mathcal{X}(\tau)$, to give a uniformly coarsely Lipschitz quasimorphism, $\psi_\tau : \mathcal{M}(\Sigma) \to \mathcal{T}(\tau)$.

By Lemma 9.1 and the subsequent remark, we get a map $\upsilon_\tau : \mathcal{T}(\tau) \to T(\tau) \subseteq \mathcal{M}(\Sigma)$, such that $\psi_\tau \circ \upsilon_\tau : \mathcal{T}(\tau) \to \mathcal{T}(\tau)$ is the identity up to bounded distance. Note that υ_τ is also a uniformly coarsely Lipschitz quasimorphism, whose image is a uniformly bounded Hausdorff distance from $T(\tau)$.

This in turn gives rise to a quasimorphism, $\omega_\tau = \upsilon_\tau \circ \psi_\tau : \mathcal{M}(\Sigma) \to T(\tau)$. It is characterised by the property that $\psi_X \circ \omega_\tau \sim \psi_X$ for all $X \in \mathcal{X}(\tau)$, or equivalently, that $\theta_Y \circ \omega_\tau \sim \theta_Y$ for all $Y \in \mathcal{X}_I(\tau)$. We note:

Lemma 9.2. *Given $r \ge 0$, there is some r' depending only on r and the parameters of the hypotheses, such that for any multicurve, τ, $T(\tau; r) \subseteq N(T(\tau); r')$.*

Proof. Let $b = \omega_\tau(a) \in T(\tau)$. By the above, we have $\theta_Y a \sim \theta_Y b$ for all $Y \in \mathcal{X}_I(\tau)$. Also $\theta_Y a \sim \theta_Y \tau \sim \theta_Y b$ for all $Y \in \mathcal{X}_T(\tau)$. Since $\mathcal{X} = \mathcal{X}_I(\tau) \cup \mathcal{X}_T(\tau)$, we see that $R_\Sigma(a, b)$ is bounded. Property (A7) "distance bound" now tells us that $a \sim b$. □

This shows that $T(\tau; r)$ is well defined up to bounded Hausdorff distance for all $r \ge r_0'$, and can be described as the set of $a \in \mathcal{M}(\Sigma)$ such that $\theta_Y a \sim \theta_Y \tau$ for all $Y \in \mathcal{X}_T(\tau)$.

Lemma 9.3. *$T(\tau)$ is uniformly quasiconvex in $\mathcal{M}(\Sigma)$.*

Proof. Suppose $a, b \in T(\tau)$ and $c \in \mathcal{M}(\Sigma)$. If $X \in \mathcal{X}_T(\tau)$, then

$$\theta_X \mu_\Sigma(a, b, c) \sim \mu_X(\theta_X a, \theta_X b, \theta_X c) \sim \mu_X(\theta_X \tau, \theta_X \tau, \theta_X c) \sim \theta_X \tau,$$

and so by Lemma 9.2, $\mu_\Sigma(a, b, c)$ is a bounded distance from $T(\tau)$. □

Lemma 9.4. *The map $\omega_\tau : \mathcal{M}(\Sigma) \to T(\tau)$ is a coarse gate map.*

Proof. Let $x \in \mathcal{M}(X)$ and $c \in T(\tau)$. If $X \in \mathcal{X}_T(\tau)$, then $\omega_\tau x \in T(\tau)$, and

$$\theta_X \mu_\Sigma(x, \omega_\tau x, c) \sim \mu_X(\theta_X x, \theta_X \omega_\tau x, \theta_X c) \sim \mu_X(\theta_X x, \theta_X \tau, \theta_X \tau) \sim \theta_X \tau \sim \theta_X \omega_\tau x.$$

If $X \in \mathcal{X}_I(\tau)$, then

$$\theta_X \mu_\Sigma(x, \omega_\tau x, c) \sim \mu_X(\theta_X x, \theta_X \omega_\tau x, \theta_X c) \sim \mu_X(\theta_X x, \theta_X x, \theta_X c) \sim \theta_X x \sim \theta_X \omega_\tau x.$$

Since $\mathcal{X} = \mathcal{X}_I(\tau) \cup \mathcal{X}_T(\tau)$, property (A7) "distance bound" tells us $\mu_\Sigma(x, \omega_\tau x, c) \sim \omega_\tau x$, as required. □

Note that, if $a, b \in \mathcal{M}(\Sigma)$, then $\omega_\tau a$ lies in a coarse interval from a to b, and $\omega_\tau b$ lies on a coarse interval from a to $\omega_\tau a$. By Lemma 6.6, ω_τ is a uniform quasimorphism (depending on $\xi(\Sigma)$). Note that the proof of Lemma 9.4 shows that $\mu_X(\theta_X a, \theta_X \omega_\tau a, \theta_X \omega_\tau b) \sim \theta_X \omega_\tau a$ (putting $x = a$ and $c = b$). Similarly, $\mu_X(\theta_X b, \theta_X \omega_\tau b, \theta_X \omega_\tau a) \sim \theta_X \omega_\tau b$. It follows that $\rho(\omega_\tau a, \omega_\tau b)$ is bounded above by a linear function of $\rho(a, b)$. We see that $\sigma_X(\omega_\tau a, \omega_\tau b)$ is bounded above by a linear function of $\sigma_X(a, b)$. In particular, if $\theta_X a \sim \theta_X b$, then $\theta_X(\omega_\tau a) \sim \theta_X(\omega_\tau b)$.

Lemma 9.5. *If τ and τ' are two multicurves which together fill Σ, then the diameter of $\omega_\tau(T(\tau'))$ in $\mathcal{M}(\Sigma)$ is finite and bounded above in terms of $\xi(\Sigma)$.*

(Here, we are assuming that we have fixed, once and for all, the constant r_0' used in defining $T(\tau)$, in terms of $\xi(\Sigma)$.

Proof. Note that if $X \in \mathcal{X}$, then either $X \in \mathcal{X}_T(\tau)$ or $X \in \mathcal{X}_T(\tau')$. Let $a, b \in T(\tau')$. If $X \in \mathcal{X}_T(\tau)$, then $\theta_X(\omega_\tau a) \sim \theta_X \tau \sim \theta_X(\omega_\tau b)$. If $X \in \mathcal{X}_T(\tau')$ then $\theta_X a \sim \theta_X b$, and by the observation preceding the lemma, $\theta_X(\omega_\tau a) \sim \theta_X(\omega_\tau b)$. It now follows by (A7) "distance bound" that $a \sim b$ as required. □

If τ, τ' fill Σ, we choose elements $\omega_\tau(\tau') \in \omega_\tau(T(\tau'))$ and $\omega_{\tau'}(\tau) \in \omega_{\tau'}(T(\tau))$. These are well defined up to bounded distance.

Now if $a \in T(\tau)$ and $b \in T(\tau')$, then $\omega_{\tau'}(\tau)$ lies in a coarse interval from a to b, and $\omega_\tau(\tau')$ lies in a coarse interval from a to $\omega_{\tau'}(\tau)$. It follows that $\rho(a, b) \asymp \rho(a, \omega_\tau(\tau')) + \rho(\omega_\tau(\tau'), \omega_{\tau'}(\tau)) + \rho(\omega_{\tau'}(\tau), b)$.

We can use this observation to prove the following.

Lemma 9.6. *Suppose that τ, τ' fill Σ, and that any pair of points of $T(\tau') \subseteq \mathcal{M}(\Sigma)$ lie a bounded distance from some uniform bi-infinite quasigeodesic in $T(\tau')$. Then there are constants $k, t \geq 0$, depending only on the constants of the hypotheses, such that if $x \in T(\tau')$ and $r \geq 0$, then there is some $y \in T(\tau')$ with $\rho(y, T(\tau)) \geq r$ and $\rho(x, y) \leq kr + t$.*

Proof. From the hypotheses, there is a uniformly quasigeodesic ray with basepoint $\omega_{\tau'}(\tau)$ and containing x. Now choose y to be a suitable point on this ray beyond x, and apply the above observation. □

For the remainder of this section, we explore these statements further in the specific case where $\mathcal{M}(\Sigma) = \mathbb{M}(\Sigma)$ is the marking graph of Σ. (Note that in this case all the parameters of $\mathbb{M}(\Sigma)$ depend only on $\xi(\Sigma)$.)

Given $k \geq 0$, let $\hat{T}(\tau; k) = \{a \in \mathbb{M}^0(\Sigma) \mid \iota(a, \tau) \leq k\}$.

Lemma 9.7. (1) *For all $k \geq 0$, there is some $r \geq 0$, depending on k and $\xi(\Sigma)$, such that $\hat{T}(\tau; k) \subseteq N(T(\tau), r)$.*

(2) *There is some $k_0 \geq 0$, depending only on $\xi(\Sigma)$, such that $T(\tau) \cap \mathbb{M}^0(\Sigma) \subseteq \hat{T}(\tau; k_0)$.*

Proof. (1) It is an elementary property of subsurface projection that if $a \in \mathbb{M}^0(\Sigma)$, $\gamma \in \mathbb{G}^0(\Sigma)$ and $X \in \mathcal{X}$ with $\gamma \prec X$ or $\gamma \pitchfork X$, then $\sigma_X(\gamma, a)$ is bounded above in terms of $\iota(\gamma, a)$. It follows that $\hat{T}(\tau; k) \subseteq T(\tau; r'')$ for some r'' depending only on r and k. We now apply Lemma 9.2.

(2) We have observed that, in the case of markings, the verification of (A10) "realisation" gives us a marking which has bounded intersection with τ (where τ is the union of all boundary curves of the set of surfaces). This was used in the construction of υ_τ and hence of ω_τ. In particular, it follows that if $a \in \mathbb{M}(\Sigma)$, then $\iota(\omega_\tau a, \tau)$ is bounded for any $a \in \mathbb{M}(\Sigma)$. Now if $a \in T(\tau) \cap \mathbb{M}^0(\Sigma)$ then (since ω_τ is a coarse gate map) $\rho(a, \omega_\tau a)$ is bounded. It follows that $\iota(a, \omega_\tau a)$ is bounded, and so $\iota(a, \tau)$ is bounded. This bound depends only on $\xi(\Sigma)$. $\square$

Definition. A *complete multicurve* is a multicurve, τ, such that each component of $\Sigma \setminus \tau$ is an $S_{0,3}$.

In other words, $\mathcal{X}_N(\tau) = \varnothing$, so $\mathcal{X}(\tau) = \mathcal{X}_A(\tau)$. It is equivalent to saying that τ has exactly $\xi(\Sigma)$ components. (It is essentially the same thing as a "pants decomposition" in other terminology.)

Suppose that τ is a complete multicurve. In this case,

$$\mathcal{X}(\tau) = \mathcal{X}_A(\tau) = \{X(\gamma) \mid \gamma \in \tau\}.$$

If $X \in \mathcal{X}(\tau)$, then $\mathbb{G}(X) = \mathbb{M}(X)$ is quasi-isometric to the real line, and so $\mathcal{T}(\tau)$ is quasi-isometric to $\mathbb{R}^\xi$. Thus, υ_τ gives rise to a quasi-isometric embedding of $\mathbb{R}^\xi$ into $\mathbb{M}(\Sigma)$, whose image is a bounded Hausdorff distance from $T(\tau)$.

We can also view this in terms of the action of $\mathrm{Map}(\Sigma)$. Let $G(\tau) \leq \mathrm{Map}(\Sigma)$ be the group generated by Dehn twists about the elements of τ. Thus, $G(\tau) \cong \mathbb{Z}^\xi$. We put the standard word metric on $G(\tau)$. Now $G(\tau)$ acts coboundedly on $\mathcal{T}(\tau)$ hence also on $T(\tau)$.

The following result, proven in [Farb, Lubotzky and Minsky 2001], is an immediate consequence, though it also follows directly from the distance formula [Masur and Minsky 2000] (given as Theorem 8.3 here).

Lemma 9.8. *Given any multicurve τ there is some $a \in \mathbb{M}(\Sigma)$ such that the map $a \mapsto ga : G(\tau) \to \mathbb{M}(\Sigma)$ is a uniform quasi-isometric embedding.*

In fact, we can take any $a \in T(\tau)$, and the orbit, $G(\tau)a$, is a uniformly bounded Hausdorff distance from $T(\tau)$. (The uniformity is somewhat spurious here, since there are only finitely many orbits of multicurves under the action of $\mathrm{Map}(\Sigma)$; though our arguments give explicit bounds.)

We will refer to a set of the form $T(\tau)$ for a complete multicurve, τ, as a *coarse Dehn twist flat* (generally regarded as defined up to a uniformly bounded Hausdorff distance).

Lemma 9.9. *There are uniform constants $k, t \geq 0$ such that if τ, τ' are complete multicurves, with $\tau \neq \tau'$, $x \in T(\tau')$ and $r \geq 0$, then there is some $y \in T(\tau')$ with $\rho(y, T(\tau)) \geq r$ and $\rho(x, y) \leq kr + t$.*

Proof. If $\tau \cap \tau' = \varnothing$, then τ, τ' fill Σ, so the result follows immediately from Lemma 9.6. (Note that any path in a Dehn twist flat lies in a uniform bi-infinite quasigeodesic.)

For the general case, let $\tau_0 = \tau \cap \tau'$. Now $T(\tau_0)$ is, up to quasi-isometry, a direct product of a euclidean space (given by Dehn twists about the elements of τ_0) and copies of $\mathbb{M}(X)$ as X ranges over the elements of $\mathcal{X}_N(\tau_0)$. Applying the above intrinsically to the restrictions of τ and τ' to any such X we deduce the general case. $\square$

10. Quasicubes

Throughout this section, we again suppose that $\mathcal{M}(\Sigma)$ satisfies the axioms (A1)–(A10) of Section 7. We refer the constants involved as the "parameters of $\mathcal{M}(\Sigma)$".

Definition. A *quasicube* in $\mathcal{M}(\Sigma)$ is an l-quasimorphism $\phi : Q \to \mathcal{M}(\Sigma)$, where Q is an n-cube.

We refer to it as an l-quasi-n-cube, if we want to specify the parameters.

In this section, we give a description of "nondegenerate" quasicubes of maximal rank. In Section 11, we will apply this to the (extended) asymptotic cone $\mathcal{M}^*(\Sigma)$.

We begin by recalling the following fact:

Lemma 10.1. *There is some $l_0 \geq 0$, depending only $\xi(\Sigma)$ such that if $X, Y \in \mathcal{X}$ and there exist $a, b, c, d \in \mathcal{M}(\Sigma)$ with $(a, b : c, d)_X \geq l_0$ and $(a, c : b, d)_Y \geq l_0$, then $X \wedge Y$.*

Here $(a, b : c, d)_X$ denotes the "crossratio"

$$(a,b:c,d)_X = \tfrac{1}{2}(\max\{\sigma_X(a,c)+\sigma_X(b,d), \sigma_X(a,d)+\sigma_X(b,c)\} - (\sigma_X(a,b)+\sigma_X(c,d)))$$

in $\mathcal{G}(X)$. Similarly for $(a, b : c, d)_Y$ in $\mathcal{G}(Y)$.

Proof. This is property (P4) in [Bowditch 2013], and was verified for the marking graph by Lemma 11.7 of that paper. As already observed (in Section 7 here) the proof there only made use of properties (A8) "bounded image" and (A9) "overlapping subsurfaces". □

Recall, from Section 7, that $\mathcal{A}_X(a, b; r) = \{Y \in \mathcal{X}(X) \mid \sigma_Y(a, b) > r\}$.

Definition. Given $a, b \in \mathcal{M}(X)$ and $r \geq 0$, we say that a, b are *weakly* (X, r)*-related* if for all $Y \in \mathcal{A}(a, b; r)$, we have $Y \preceq X$.

Intuitively, we can think of a, b as being "close outside X". More specifically, if $Z \in \mathcal{X}$, with $Z \wedge X$, then $\mathcal{A}(a, b; r) \cap \mathcal{X}(Z) = \varnothing$, so by (A7) "distance bound", we see that $\rho_Z(a, b)$ is bounded.

Definition. We say that a, b are (X, r)*-related* if they are weakly (X, r)-related and $\rho(a, T(\partial_\Sigma X)) \leq r$ and $\rho(b, T(\partial_\Sigma X)) \leq r$.

We will often suppress mention of r where a choice (ultimately depending on the parameters of $\mathcal{M}(\Sigma)$) is clear from context, and simply refer to a, b as being "(weakly) X-related".

Note that this property is "median convex" in the sense that if $c \in \mathcal{M}(\Sigma)$, with $\rho(\mu(a, b, c), c) \leq l$, then c is (X, r')-related to a and to b, where r' depends only on r, l and the parameters of the hypotheses.

Lemma 10.2. *If $a, b \in \mathcal{M}(\Sigma)$ are (X, r)-related, then $\rho(a, b)$ agrees with $\rho(a, b)$ up to linear bounds depending only on r and $\xi(\Sigma)$.*

Proof. The fact that $\psi_X : \mathcal{M}(\Sigma) \to \mathcal{M}(X)$ is uniformly coarsely Lipschitz immediately gives a linear upper bound for $\rho_X(a, b)$. For the other direction, set $\tau = \partial_\Sigma X$. Then, up to bounded distance, $\psi_\tau a$ and $\psi_\tau b$ differ only in the X-coordinate. Since υ_τ is uniformly coarsely Lipschitz, we have that $\rho(\omega_\tau a, \omega_\tau b) = \rho(\upsilon_\tau \psi_\tau a, \upsilon_\tau \psi_\tau b)$ is linearly bounded above by $\rho_X x(\psi_X a, \psi_X b) = \rho_X(a, b)$. By assumption $\rho(a, T(\tau))$ and $\rho(b, T(\tau))$ are bounded. By Lemma 9.4, since ω_τ is a coarse gate map to $T(\tau)$, it follows that $\rho(a, \omega_\tau a)$ and $\rho(b, \omega_\tau b)$ are bounded. This bounds $\rho(a, b)$, as required. □

Note that by median convexity, we see also that $\rho_X(x, y) \asymp \rho(x, y)$ for any $x, y \in \mathcal{M}(\Sigma)$ with $\mu(a, b, x) \sim x$ and $\mu(a, b, y) \sim y$.

Here is a criterion which implies that two elements of $\mathcal{M}(\Sigma)$ are X-related:

Lemma 10.3. *There is a constant $r_1 \geq 0$ depending only on $\xi(\Sigma)$ with the following property. Suppose $r_2 \geq r_1$, and that $a, b \in \mathcal{M}(\Sigma)$ and $X \in \mathcal{X}$. Suppose that for all $Z \in \mathcal{A}(a, b; r_2)$ we have $Z \preceq X$. Suppose moreover that whenever γ is a curve with $\gamma \pitchfork X$, there is some $Y \in \mathcal{A}_\Sigma(a, b; r_1)$ with $\gamma \pitchfork Y$ and $Y \preceq X$. Then a, b are (X, r)-related for some r depending only on r_2 and the parameters of $\mathcal{M}(\Sigma)$.*

Proof. By assumption, a, b are weakly (X, r)-related, so we just need to check that $\rho(a, T(\tau))$ and $\rho(b, T(\tau))$ are bounded, where $\tau = \partial_\Sigma(X)$.

By Lemma 9.2, it is enough to check that $\sigma_U(a, \tau)$ and $\sigma_U(b, \tau)$ are bounded for all $U \in \mathcal{X}_T(\tau)$. Now if $U \in \mathcal{X}_T(\tau)$, then U contains or crosses some boundary curve of X, and so $U \pitchfork X$ or $X \prec U$. Either way, U will contain a curve, γ, with $\gamma \pitchfork X$. By hypothesis, there is some $Y \preceq X$, with $Y \pitchfork \gamma$ and $Y \in \mathcal{A}_\Sigma(a, b; r_1)$. Note that either $Y \prec U$ or $Y \pitchfork U$.

Now $\sigma_Y(a, b) > r_1$. Also, since U is not contained in X, it does not lie in $\mathcal{A}_\Sigma(a, b; r_2)$, i.e., $\sigma_U(a, b) \leq r_2$. Suppose first that $Y \prec U$. If r_1 is greater than r_0, the constant of (A8) "bounded image", then it follows $\langle \theta_U a, \theta_U b{:}\theta_U Y \rangle \leq r_0$, and so, by the definition of Gromov product, $\sigma_U(a, Y) + \sigma_U(b, Y) \leq 2(r_2 + r_0)$. Suppose instead that $Y \pitchfork U$. If r_1 is bigger than twice the constant, r_0, of (A9) "overlapping subsurfaces", then without loss of generality (swapping a and b), we have $\sigma_Y(a, U) > r_0$, so by (A9) we must have $\sigma_U(a, Y) \leq r_0$. Since $\sigma_U(a, b) \leq r_2$, it follows that $\sigma_U(b, Y) \leq r_2 + r_0$. Thus, in all cases, we have shown that $\sigma_U(a, Y)$ and $\sigma_U(b, Y)$ are bounded.

But now, $Y \preceq X$, so $Y \wedge \tau$. Thus, by (A5) "disjoint projection", we have that $\sigma_U(Y, \tau)$ is bounded. We deduce that $\sigma_U(a, \tau)$ and $\sigma_U(b, \tau)$ are bounded for all $U \in \mathcal{X}_T(\tau)$ as claimed. □

We now move on to consider quasicubes.

Suppose that $Q = \{-1, 1\}^n$ is an n-cube. By an *i-th side* of Q, we mean an unordered pair, $c, d \in Q$, which differ precisely in their i-th coordinates. Note that any two i-th sides are parallel in the median sense. If $a, b \in Q$, we can speak of the sides of Q *crossed* by a, b, that is those which (up to parallelism) correspond to the coordinates for which a, b differ. (Note that the walls of Q are in bijective correspondence with the parallel classes of sides.)

Suppose that $\phi : Q \to \mathcal{M}(\Sigma)$ is an l-quasimorphism. If c, d and c', d' are both i-th sides of Q, then $\rho(\phi c, \phi d) \asymp \rho(\phi c', \phi d')$. (Since $\mu(\phi c, \phi d, \phi c') \sim \phi c$ and $\mu(\phi c, \phi d, \phi d') \sim \phi d$, we get a linear upper bound for $\rho(\phi c, \phi d)$, and the lower bound follows symmetrically.) We will write $s_i = \min \rho(\phi c, \phi d)$ as c, d ranges over all i-th sides. Thus $\rho(\phi c', \phi d') \asymp s_i$ for any other i-th sides, c', d'. We also note that for all $X \in \mathcal{X}$, we have $\sigma_X(\phi c, \phi d) \asymp \sigma_X(\phi c', \phi d')$ and $\rho_X(\phi c, \phi d) \asymp \rho_X(\phi c', \phi d')$ (similarly, since $\theta_X \circ \phi$ and $\psi_X \circ \phi$ are quasimorphisms to $\mathcal{G}(X)$ and $\mathcal{M}(X)$ respectively). Here, the linear bounds depend implicitly on l.

If $a, b \in Q$, then a repeated application of Lemma 6.1 shows $\rho(\phi a, \phi b) \asymp \sum_i s_i$, where the sum is taken over all sides of Q crossed by a, b.

Lemma 10.4. *Let $\phi : Q \to \mathcal{M}$ be an l-quasicube. There is some $k_0 \geq 0$, depending only on h and the parameters of the hypotheses, such that if $X, Y \in \mathcal{X}$ with*

$\sigma_X(\phi c, \phi d) \geq k_0$ and $\sigma_Y(\phi c', \phi d') \geq k_0$, where c, d and c', d' are respectively i-th and j-th sides of Q, then either $i = j$ or $X \wedge Y$.

Proof. This is an immediate consequence of Lemma 10.1 above. □

It follows from Lemma 10.4 that for each i, there is a (possibly empty, possibly disconnected) subsurface, Y_i, of Σ, which contains all $X \in \mathcal{X}$, for which $\sigma_X(\phi c, \phi d) \geq k_0$ for any i-th side, c, d, of Q. (Here, we are using the term "subsurface" to mean a disjoint union of essential subsurfaces as defined in Section 7.) We can also take Y_i to be minimal with this property. To ensure that each Y_i must be nonempty, we will assume that ϕ is *nondegenerate*, that is, for each side, c, d, $\mathcal{A}_\Sigma(c, d; r) \neq \emptyset$, for a fixed $r \geq k_0$. Note that, by (A7) "distance bound", this is implied by placing a suitable lower bound on $\min\{s_i \mid 1 \leq i \leq n\}$. We will also take r to be at least the constant r_1 featuring in Lemma 10.3.

Recall that $\xi(\Sigma)$ is the maximal number of disjoint and distinct (non-$S_{0,3}$) subsurfaces we can embed in Σ. Thus, if $n = \xi(\Sigma)$, we see that if ϕ is nondegenerate, then each Y_i is connected and is either an annulus, or has complexity-1 (that is a $S_{0,4}$ or $S_{1,1}$).

Definition. We say that a multicurve τ is *big* if each component of $\Sigma \setminus \tau$ is a $S_{0,3}$, $S_{0,4}$ or $S_{1,1}$.

In this case, the set of all relative boundary components of all the Y_i is a big multicurve, τ, such that $\mathcal{X}(\tau)$ is precisely the set of Y_i.

Lemma 10.5. *Suppose that Q is a ξ-cube, and that $\phi : Q \to \mathcal{M}(\Sigma)$ is a nondegenerate quasimorphism. Then there is a big multicurve, $\tau \subseteq \Sigma$, such that we can write $\mathcal{X}(\tau) = \{Y_1, \ldots, Y_\xi\}$, so that if c, d is any i-th side of Q, then $\phi c, \phi d$ are Y_i-related.*

(We remark that it follows that $\phi(Q)$ lies in a bounded neighbourhood of $T(\tau)$.)

Proof. We construct disjoint surfaces Y_i as above, and as already observed, the set of these Y_i is precisely $\mathcal{X}(\tau)$ for a big multicurve, τ. Recall that for all $X \in \mathcal{X}$, if c, d and c', d' are i-th sides of Q, the $\sigma_X(\phi c, \phi d) \asymp \sigma_X(\phi c', \phi d')$. Let r_1 be the constant of Lemma 10.3. If the nondegeneracy constant is sufficiently large, then $\mathcal{A}(\phi c, \phi d; r) \neq \emptyset$. So if $r \geq \max\{r_1, k_0\}$, the subsurfaces of $\mathcal{A}(\phi c, \phi d; r)$ fill Y_i, so $\phi c, \phi d, Y_i$ satisfy the hypotheses of Lemma 10.3. We see that $\phi c, \phi d$ are Y_i-related as claimed □

Note that a big multicurve τ satisfying the conclusion of Lemma 10.5 might not be unique. For example, if $\gamma \in \tau$ bounds an $S_{0,3}$ component of $\Sigma \setminus \tau$ on both sides (perhaps the same $S_{0,3}$), then we can remove it, and the conclusion will still hold. However, this is essentially the only ambiguity that can arise.

We will also need:

Lemma 10.6. *Suppose that Q is a ξ-cube, and that $\phi : Q \to \mathcal{M}(\Sigma)$ is a non-degenerate quasimorphism, and that c, d is an i-th side of Q. Let Y_i be as given by Lemma 10.5. Suppose that $x, y \in \mathcal{M}(\Sigma)$ with $\mu(\phi c, \phi d, x) \sim x$ and $\mu(\phi c, \phi d, y) \sim y$. Then $\rho(x, y) \asymp \rho_{Y_i}(x, y)$, where the additive bounds depend only on l, n and the parameters of $\mathcal{M}(\Sigma)$.*

Proof. By Lemma 10.4, and ϕc, ϕd are Y_i-related, and so therefore are x, y (with suitable constants). The statement then follows by Lemma 10.2 and the subsequent observation. □

11. The asymptotic cone of $\mathcal{M}(\Sigma)$

As in Section 5, let $\mathcal{M}^*(\Sigma)$ be the extended asymptotic cone of a space $\mathcal{M}(\Sigma)$ satisfying the hypotheses (A1)–(A10).

Let $\mathcal{Z}$ be a countable set with a nonprincipal ultrafilter, as in Section 5. Let $\mathcal{U}\mathbb{G} = \mathcal{U}\mathbb{G}(\Sigma)$ be the ultrapower of $\mathbb{G}(\Sigma)$. This is a graph with vertex set $\mathcal{U}\mathbb{G}^0$. Note that the intersection number, ι, extends to a map $\mathcal{U}\iota : (\mathcal{U}\mathbb{G}^0)^2 \to \mathcal{U}\mathbb{N}$. We also have an ultrapower, $\mathcal{U}\mathcal{X} = \mathcal{U}\mathcal{X}_A \sqcup \mathcal{U}\mathcal{X}_N$. There is a natural bijection between $\mathcal{U}\mathcal{X}_A$ and $\mathcal{U}\mathbb{G}^0$. (Here, $\mathbb{G}^0$ is playing its role as an indexing set.)

We can extend the notation introduced in Section 5. For example, if $X, Y \in \mathcal{U}\mathcal{X}$, we write $X \wedge Y$ to mean that $X_\zeta \wedge Y_\zeta$ almost always. We similarly define $X \prec Y$ and $X \pitchfork Y$. Since there are only finitely many possibilities (in fact, five), we have the following pentachotomy: if $X, Y \in \mathcal{X}$, exactly one of $X = Y$, $X \wedge Y$, $X \prec Y$, $Y \prec X$ or $X \pitchfork Y$ must hold (exactly as in Section 7).

Note that $\mathcal{U}\,\mathrm{Map}(\Sigma)$ acts on both $\mathcal{U}\mathbb{G}$ and $\mathcal{U}\mathcal{X}$ with finite quotient.

Terminology. In this section, we refer to an element of $\mathcal{U}\mathbb{G}^0$ as a *curve* and an element of $\mathbb{G}^0 \subseteq \mathcal{U}\mathbb{G}^0$ as *standard curve*. We similarly refer to "subsurfaces", "standard subsurfaces" etc.

As observed in Section 5, two standard objects lie in the same $\mathcal{U}\,\mathrm{Map}(\Sigma)$-orbit, then they lie in the same $\mathrm{Map}(\Sigma)$-orbit.

Moreover, any configuration of curves and surfaces of bounded complexity can be assumed standard up to the action of the mapping class group. One way to express this is as follows.

Lemma 11.1. *Suppose that $a \subseteq \mathcal{U}\mathbb{G}^0$ and $\mathcal{U}\iota(a) \in \mathbb{N}$, then there is some $g \in \mathcal{U}\,\mathrm{Map}(\Sigma)$ with $ga \subseteq \mathbb{G}^0$.*

Proof. By hypothesis, $\iota(a_\zeta)$ is almost always constant. Therefore, we can find $g_\zeta \in \mathrm{Map}(\Sigma)$ such that $g_\zeta a_\zeta \subseteq \mathbb{G}^0(\Sigma)$ lies is one of only finitely many possible subsets of $\mathbb{G}^0(\Sigma)$. Therefore, $g_\zeta a_\zeta$ is almost always constant, that is, ga is standard, where g is the limit of $(g_\zeta)_\zeta$. □

Note that this applies, for example, to multicurves, or to collections of pairwise disjoint subsurfaces of Σ. In particular, it makes sense to refer to the topological type of a subsurface; for example, that it is an $S_{1,1}$ or an $S_{0,4}$ (up to the action of $\mathcal{U}\,\mathrm{Map}(\Sigma)$). We can also refer to boundary curves of a surface, or say that a collection of curves fill a subsurface, etc.

If τ is a multicurve, we can define $\mathcal{U}\mathcal{X}(\tau) \subseteq \mathcal{U}\mathcal{X}$ as in Section 9. (It is the limit of the sets $\mathcal{X}(\tau_\zeta)$.)

In what follows we deal mostly with extended asymptotic cones. This seems more natural in this context than restricting to the asymptotic cone, though most of the discussion would apply equally well in both situations.

We assume that $\mathcal{M}(\Sigma)$ and $\mathcal{G}(\Sigma)$ are spaces satisfying properties (A1)–(A10) of Section 7.

Let $\boldsymbol{t} \in \mathcal{U}\mathbb{R}$ be a positive infinitesimal. Rescaling as in Section 5 we get extended asymptotic cones, $\mathcal{M}^* = \mathcal{M}^*(\Sigma)$ and $\mathcal{G}^* = \mathcal{G}^*(\Sigma)$ of $\mathcal{M}(\Sigma)$ and $\mathcal{G}(\Sigma)$, respectively. (We topologise them as the disjoint union of their components.) We write ρ^*, σ^*, respectively, for the limiting nonstandard metrics. The asymptotic cones, $(\mathcal{M}^\infty(\Sigma), \rho^\infty)$ and $(\mathcal{G}^\infty(\Sigma), \sigma^\infty)$ are complete metric spaces. In fact, since $\mathcal{G}(\Sigma)$ is hyperbolic, $\mathcal{G}^*(\Sigma)$ is an $\mathbb{R}^*$-tree, and $\mathcal{G}^\infty(\Sigma)$ is an $\mathbb{R}$-tree. The map $\chi : \mathcal{M}(\Sigma) \to \mathcal{G}(\Sigma)$ is coarsely Lipschitz, and (after the rescaling) gives rise to a Lipschitz map $\chi^* : \mathcal{M}(\Sigma) \to \mathcal{G}(\Sigma)$.

The coarse median μ on $\mathcal{M}(\Sigma)$ gives rise to a median μ^* on $\mathcal{M}^*$, and restricts to a median, μ^∞ on the asymptotic cone, $\mathcal{M}^\infty$. By Theorems 6.8 and 6.9, ρ^∞ is bi-Lipschitz equivalent to a median metric inducing the same median structure. The construction is not canonical, but we will write ρ_M^∞ for some choice of such median metric. We similarly define ρ_M^* bi-Lipschitz equivalent to ρ^* on $\mathcal{M}^*$.

Suppose that $X \in \mathcal{U}\mathcal{X}$. The spaces $\mathcal{G}(X_\zeta)$ give rise to an extended asymptotic cone, denoted $\mathcal{G}^*(X)$ which is an $\mathbb{R}^*$-tree. The maps θ_{X_ζ} are uniformly coarsely Lipschitz, and give rise to a Lipschitz homomorphism, $\theta_X^* : \mathcal{M}^*(X) \to \mathcal{G}^*(X)$.

Similarly, we have a limit $\mathcal{M}^*(X)$ of the spaces $\mathcal{M}(X_\zeta)$. This has a median μ^* arising from the coarse medians μ_ζ and we again get a topological median algebra. We similarly have limiting Lipschitz homomorphisms $\psi_X^* : \mathcal{M}^*(\Sigma) \to \mathcal{M}^*(X)$. In fact, as observed above, up to the action of $\mathcal{U}\,\mathrm{Map}(\Sigma)$, we could take X to be standard, and so $\mathcal{M}^*(X)$ is isomorphic to the space defined intrinsically on a surface of this topological type.

If $X, Y \in \mathcal{U}\mathcal{X}$, with $X \preceq Y$, then we have a limiting map, $\psi_{YX}^* : \mathcal{M}^*(X) \to \mathcal{M}^*(Y)$, with $\psi_{YX}^* \circ \psi_X^* = \psi_Y^*$. We will generally abbreviate ψ_{XY}^* to ψ_X^*, when the domain is clear from context.

Note that if $\gamma \in \mathcal{U}\mathbb{G}^0$ and $X \in \mathcal{U}\mathcal{X}$ with $\gamma \preceq X$ or $\gamma \pitchfork X$, we have a well defined subsurface projection, $\theta_X^*(\gamma) \in \mathcal{G}^*(X)$. Similarly, if $X, Y \in \mathcal{U}\mathcal{X}$, with $Y \prec X$ or $Y \pitchfork X$, we can define $\theta_X^*(Y) \in \mathcal{G}^*(X)$.

We also note that if $\gamma \in \mathcal{U}\mathbb{G}^0$, then we can write $\mathcal{M}^*(\gamma) = \mathcal{G}^*(\gamma) = \mathcal{G}^*(X)$ where $X = X(\gamma)$ is an annular neighbourhood of γ.

Suppose that $\tau \subseteq \mathcal{U}\mathbb{G}^0$ is a multicurve. Let $T^*(\tau) \subseteq \mathcal{M}^*(\Sigma)$ be the limit of the subsets $T(\tau_\zeta) \subseteq \mathcal{M}(\Sigma)$. This is a closed subset of $\mathcal{M}^*(\Sigma)$. (Note that it is also the limit of the sets $T(\tau_\zeta; r)$ for any sufficiently large $r \in [0, \infty)$.)

We can describe the structure of $T^*(\tau)$ as follows.

Let $\mathcal{U}\mathcal{X}(\tau) \subseteq \mathcal{U}\mathcal{X}$ be the ultraproduct of the $\mathcal{X}(\tau_\zeta)$. (By Lemma 11.1, this is finite, and standard up to the action of $\mathcal{U}\,\mathrm{Map}(\Sigma)$.) Let $\mathcal{T}^*(\tau)$ be the direct product of the spaces $\mathcal{M}^*(X)$ for $X \in \mathcal{U}\mathcal{X}(\tau)$ in the l^1 extended metric. This is the same as the extended asymptotic cone of the spaces $\mathcal{T}(\tau_\zeta)$.

Recall that in Section 9, we defined maps $\psi_{\tau_\zeta} : \mathcal{M}(\Sigma) \to \mathcal{T}(\tau_\zeta)$, $\upsilon_{\tau_\zeta} : \mathcal{T}(\tau_\zeta) \to T(\tau_\zeta)$ and $\omega_{\tau_\zeta} = \upsilon_{\tau_\zeta} \circ \psi_{\tau_\zeta} : \mathcal{M}(\Sigma) \to \mathcal{T}(\tau_\zeta)$. These are all uniformly coarsely Lipschitz quasimorphisms, and so give rise to maps, $\psi^*_\tau : \mathcal{M}^*(\Sigma) \to \mathcal{T}^*(\tau)$, $\upsilon^*_\tau : \mathcal{T}^*(\tau) \to T(\tau)$ and $\omega^*_\tau = \upsilon^*_\tau \circ \psi^*_\tau : \mathcal{M}^*(\Sigma) \to \mathcal{T}^*(\tau)$. In fact, from Lemma 9.4, we see that $\omega^*_\tau : \mathcal{M}^*(\Sigma) \to T^*(\tau)$ is a gate map.

It follows that $T^*(\tau)$ is convex, and that ω^*_τ is the unique gate map to $T^*(\tau)$. Note also that if $\gamma \in \mathcal{U}\mathbb{G}^0(\Sigma)$ with $\tau \pitchfork \gamma$ and $a \in T^*(\tau)$, then $\theta^*_\gamma(a) = \theta^*_\gamma(\tau)$.

Given a subset, $S \subseteq \mathcal{M}^*(\Sigma)$, write

$$C(S) = \{X \in \mathcal{U}\mathcal{X} \mid \theta^*_X|S \text{ is injective}\},$$

and

$$D(S) = \{X \in \mathcal{U}\mathcal{X} \mid \psi^*_X|S \text{ is injective}\}.$$

Clearly $C(S) \subseteq D(S)$. We write $D^0(S) = C(S) \cap \mathcal{U}\mathcal{X}_A = D(S) \cap \mathcal{U}\mathcal{X}_A$, which we can identify as a subset of $\mathcal{U}\mathbb{G}^0$.

Given $a, b \in \mathcal{M}^*(\Sigma)$, write $C(a, b) = C([a, b])$, $D(a, b) = D([a, b])$ and $D^0(a, b) = D^0([a, b])$.

We also write $A(a, b) = C(\{a, b\}) = \{X \in \mathcal{U}\mathcal{X} \mid \theta^*_X a \neq \theta^*_X b\}$. Clearly $C(a, b) \subseteq A(a, b)$.

Lemma 11.2. *Suppose that $a, b, c, d \in \mathcal{M}^*(\Sigma)$ are all distinct. Suppose $X, Y \in \mathcal{U}\mathcal{X}$ $c \in [a, d]$, $b \in [a, c]$, $X \in A(a, b) \cap A(c, d)$ and $Y \in A(b, c)$, then either $X = Y$ or $X \wedge Y$.*

Proof. Suppose first, for contradiction, that $X \pitchfork Y$. (This is the case that is actually of interest to us.) Let $a_\zeta, b_\zeta, c_\zeta, d_\zeta \in \mathcal{M}(X_\zeta)$ be sequences converging to $a, b, c, d \in \mathcal{M}^*(\Sigma)$. We can suppose that $\rho_{X_\zeta}(c_\zeta, \mu(a_\zeta, c_\zeta, d_\zeta))$ and $\rho_{X_\zeta}(b_\zeta, \mu(a_\zeta, b_\zeta, c_\zeta))$ are bounded (after replacing c_ζ by $\mu(a_\zeta, c_\zeta, d_\zeta)$) and then b_ζ by $\mu(a_\zeta, b_\zeta, c_\zeta)$)). Now $\sigma_{X_\zeta}(a_\zeta, b_\zeta) \to \infty$ (since $\theta_{X_\zeta} a_\zeta \to \theta^*_X a$ and $\theta_{X_\zeta} b_\zeta \to \theta^*_X b$, which by hypothesis are distinct). In particular, $\sigma_{X_\zeta}(a_\zeta, b_\zeta)$ is almost always greater than $2r_0$, where r_0 is the constant of property (A9) "overlapping subsurfaces". Thus, $\theta_{X_\zeta} Y_\zeta$ must be at distance greater than r_0 from either $\theta_{X_\zeta} a_\zeta$ or $\theta_{X_\zeta} b_\zeta$, and so by (A9), $\theta_{Y_\zeta} X_\zeta$ is within a distance r_0 from either $\theta_{Y_\zeta} a_\zeta$ or $\theta_{Y_\zeta} b_\zeta$. Similarly, $\theta_{Y_\zeta} X_\zeta$ is also almost always within distance r_0 of either $\theta_{Y_\zeta} c_\zeta$ or $\theta_{Y_\zeta} d_\zeta$. But $\mathcal{G}(Y_\zeta)$ is uniformly hyperbolic, and θ_{X_ζ} is a median quasimorphism. Therefore, up to bounded distance, $\theta_{Y_\zeta} b_\zeta$ and $\theta_{Y_\zeta} c_\zeta$ lie on a geodesic from $\theta_{Y_\zeta} a_\zeta$ to $\theta_{Y_\zeta} d_\zeta$, and occur in this order. Therefore, whichever of the above possibilities arises, we see that $\sigma_{Y_\zeta}(b_\zeta, c_\zeta)$ is bounded, and so $\sigma^*_Y(b, c) = 0$. That is, $\theta^*_X b = \theta^*_X c$, so $Y \notin A(b, c)$.

After swapping the roles of X and Y if necessary, we also need the to rule out the possibility that $X \prec Y$. If that were the case, we could derive a similar contradiction using property (A8) "bounded image". Briefly, if $\sigma_{X_\zeta}(a_\zeta, b_\zeta)$ is large then $\theta_{Y_\zeta} X_\zeta$ must lie close to any geodesic in $\mathcal{G}(Y_\zeta)$ from $\theta_{Y_\zeta} a_\zeta$ to $\theta_{Y_\zeta} b_\zeta$. Similarly, $\theta_{Y_\zeta} X_\zeta$ lies close to any geodesic from $\theta_{Y_\zeta} c_\zeta$ to $\theta_{Y_\zeta} d_\zeta$. We again get that $\sigma_{Y_\zeta}(b_\zeta, c_\zeta)$ is bounded, and so derive a contradiction. (We omit details, since we will not need this case in this paper.) $\square$

We say that a subset O of $\mathcal{M}^*(\Sigma)$ is *monotone* if it admits a total order $<$ such that if $x < y < z$ in O then $y \in [x, z]$.

Recall that, $C(O)$ is the set of $X \in \mathcal{UX}$ such that $\theta^*_X | O : O \to \mathcal{G}^*(X)$ is injective. The following is an immediate corollary of Lemma 11.2.

Corollary 11.3. *If $O \subseteq \mathcal{M}^*(\Sigma)$ is monotone, $|O| \geq 4$, and $X, Y \in C(O)$, then either $X = Y$, or $X \wedge Y$.*

(In fact, the only information we really need from Lemma 11.2 and Corollary 11.3 is that X and Y cannot cross.)

Note, in particular, this applies if $O \subseteq \mathcal{M}^*(\Sigma)$ is a convex subset order-isomorphic to a totally ordered set.

In particular, Corollary 11.3 tells us that, if $|O| \geq 4$, then $C^0(O)$ is a multicurve.

Note, in particular, this applies if $O \subseteq \mathcal{M}^*(\Sigma)$ is a nontrivial convex subset order-isomorphic to a totally ordered set.

12. Cubes in $\mathcal{M}^*(\Sigma)$

Let $Q \subseteq \mathcal{M}^*(\Sigma)$ be an n-cube. If c, d and c', d' are both i-th sides of Q, then the intervals $[c, d]$ and $[c', d']$ are parallel. That is, the maps $x \mapsto \mu(c', d', x)$ and $x \mapsto \mu(c, d, x)$ are inverse median isomorphisms between $[c, d]$ and $[c', d']$. Now if $x, y \in [c, d]$, let $x' = \mu(c', d', x)$ and $y' = \mu(c', d', y)$. Given $X \in \mathcal{UX}$, and $\theta^*_X x = \theta^*_X y$, then $\theta^*_X x' = \theta^*_X y'$. We see that if $\theta^*_X | [c', d']$ is injective, then so is $\theta^*_X | [c, d]$ and conversely by symmetry. The same applies to ψ^*_X. Thus $C([c, d]) = C([c', d'])$, so we can write this as $C_i(Q)$. We similarly write $D_i(Q) = D([c, d]) = D([c', d'])$. We write $D^0_i(Q) = D_i(Q) \cap \mathcal{U}\mathbb{G}^0(\Sigma) = C_i(Q) \cap \mathcal{U}\mathbb{G}^0(\Sigma)$, which we identify with the set of curves $\gamma \in \mathcal{U}\mathbb{G}^0(\Sigma)$ such that $\theta^*_\gamma | [c, d]$ is injective.

Suppose now that $n = \xi$. In this case, if c, d is any side of Q, then $[c, d]$ is a rank-1 median algebra (a totally ordered set). We refer to $[c, d]$ as a *face* of the convex hull, hull(Q), of Q. In fact, hull(Q) is a median direct product of its faces. If $n = \xi$, then each of the faces has rank 1. Such a face is linearly ordered, and so it is isometric in ρ^*_M to an interval in $\mathbb{R}^*$ (via the map $x \mapsto \rho^*(c, x)$).

Applying Corollary 11.3, we immediately get:

Lemma 12.1. *If $X, Y \in C_i(Q)$ for any i, then either $X = Y$ or $X \wedge Y$.*

Recall that by Lemma 6.3, there are uniform quasimorphisms, $\phi_\zeta : Q \to \mathcal{M}(\Sigma)$, such that $\phi_\zeta x \to x$ for all $x \in Q$. Note that, necessarily, we have that ϕ_ζ is nondegenerate for almost all ζ.

Lemma 12.2. *If $X \in C_i(Q)$ and $Y \in C_j(Q)$, then either $i = j$ or $X \wedge Y$.*

Proof. Let $\phi_\zeta : Q \to \mathcal{M}(\Sigma)$ be $\mathcal{Z}$-sequence of uniform quasimorphisms as given by Lemma 6.7, with $\phi_\zeta x \to x$ for all $x \in Q$. Let c, d and c', d' be i-th and j-th sides of Q, respectively. Then $\sigma_{X_\zeta}(\phi_\zeta c, \phi_\zeta d) \to \infty$ and $\sigma_{Y_\zeta}(\phi_\zeta c', \phi_\zeta d') \to \infty$, and so for almost all ζ, $\sigma_{X_\zeta}(\phi_\zeta c, \phi_\zeta d) \geq k_0$ and $\sigma_{Y_\zeta}(\phi_\zeta c', \phi_\zeta d') \geq k_0$, where k_0 is the constant of Lemma 10.4. If $i \neq j$, then by Lemma 10.4, $X_\zeta \wedge Y_\zeta$, so it follows that $X \wedge Y$. □

Given that $D^0_i(Q) \subseteq C_i(Q)$, we see that if $\alpha \in D^0_i(Q)$ and $\beta \in D^0_j(Q)$, then either $\alpha = \beta$ or $\alpha \wedge \beta$, and in the former case, $i = j$.

Lemma 12.3. *Suppose that $\gamma \in D^0_i(Q)$ and that $X \in D_j(Q)$ is a complexity-*1 *subsurface* (*i.e., an $S_{0,4}$ or $S_{1,1}$*). *If $\gamma \prec X$, then $i = j$.*

Proof. Let k_0 be the constant of Lemma 10.4. Let ϕ_ζ, c, d, c', d' be as in the proof of Lemma 12.2. Now $\sigma_{\gamma_\zeta}(\phi_\zeta c, \phi_\zeta d) \to \infty$, and $\rho_{X_\zeta}(\phi_\zeta c', \phi_\zeta d') \to \infty$. Thus we have the following for almost all ζ. First, $\sigma_{\gamma_\zeta}(\phi_\zeta c, \phi_\zeta d) \geq k_0$. Second, using (A7) "distance bound", there is some $Y_\zeta \preceq X_\zeta$ with $\sigma_{Y_\zeta}(\phi_\zeta c, \phi_\zeta d) \geq k_0$. Third, $\gamma_\zeta \prec X_\zeta$ and X_ζ has complexity 1. It follows that either $\gamma_\zeta \preceq Y_\zeta$ or $\gamma_\zeta \pitchfork Y_\zeta$. But then, by Lemma 10.4, we must have $i = j$. □

Lemma 12.4. *Let $Q \subseteq \mathcal{M}^*(\Sigma)$ be a ξ-cube. Then there is a big multicurve τ such that we can write $\mathcal{UX}(\tau) = \{Y_1, \ldots, Y_\xi\}$ with the Y_i all distinct, and with $Y_i \in D_i(Q)$.*

Proof. Let $\phi_\zeta : Q \to \mathcal{M}^*(\Sigma)$ be uniform quasimorphisms as in the two previous proofs. Let τ_ζ be the standard big multicurve given by Lemma 10.5, and write $\mathcal{X}(\tau_\zeta) = \{Y_{1,\zeta}, \ldots, Y_{\xi,\zeta}\}$. Let c, d be an i-th side of Q. Then $\phi_\zeta c$ and $\phi_\zeta d$ are $Y_{i,\zeta}$-related. Write $c_\zeta = \phi_\zeta c$ and $d_\zeta = \phi_\zeta d$. Let τ be the limit of $(\tau_\zeta)_\zeta$, and let Y_i be the limit of $(Y_{i,\zeta})_\zeta$. Thus $\mathcal{UX}(\tau) = \{Y_1, \ldots, Y_\xi\}$.

It remains to show that $Y_i \in D_i(Q)$. Suppose that $x, y \in [c, d]$. Let $x_\zeta, y_\zeta \in \mathcal{M}(\Sigma)$ with $x_\zeta \to x$ and $y_\zeta \to y$. After replacing x_ζ by $\mu(c_\zeta, d_\zeta, x_\zeta)$ and y_ζ by $\mu(c_\zeta, d_\zeta, y_\zeta)$, we can assume that $\mu(c_\zeta, d_\zeta, x_\zeta) \sim x_\zeta$ and $\mu(c_\zeta, d_\zeta, y_\zeta) \sim y_\zeta$. By Lemmas 10.5, $\phi_\zeta c, \phi_\zeta d$ are X-related. By Lemma 10.2 and subsequent remarks, it follows that $\rho(x_\zeta, y_\zeta) \asymp \rho_{Y_{i,\zeta}}(x_\zeta, y_\zeta)$. But $\rho(x_\zeta, y_\zeta) \to \rho(x, y)$ and $\rho_{Y_{i,\zeta}}(x_\zeta, y_\zeta) \to$

$\rho_{Y_i}(x, y)$, and so $\rho(x, y)$ and $\rho_{Y_i}(x, y)$ are bi-Lipschitz related. In particular, if $\psi^*_{Y_i} x = \psi^*_{Y_i} y$, then $\rho_{Y_i}(x, y) = 0$, so $\rho(x, y) = 0$, so $x = y$. In other words, this shows that $\psi^*_{Y_i}|[c, d]$ is injective, so $Y_i \in D_i(Q)$ as claimed. $\square$

Let I be the set of i such that $Y_i = X(\gamma_i)$ is an annulus. Thus, $\gamma_i \in D^0_i(Q)$, and $\tau = \{\gamma_i \mid i \in I\}$. By Lemmas 12.2 and 12.4, we see that, in fact, $D^0_i(Q) = \{\gamma_i\}$. Moreover, if $D^0_j(Q) \neq \varnothing$ for some $j \notin I$, then using Corollary 11.3 and Lemmas 12.2 and 12.3, we again see that $D^0_j(Q)$ consists of a single curve γ_j, with $\gamma_j \prec Y_j$. Now write $I(Q) = \{i \mid D^0_i(Q) \neq \varnothing\}$, and let $\tau(Q) = \{\gamma_i \mid i \in I(Q)\}$. We see that $\tau(Q)$ is a big multicurve containing τ and that it also satisfies the conclusion of Lemma 12.4 since $\mathcal{UX}_N(\tau(Q)) \subseteq \mathcal{UX}_N(\tau)$. (Therefore retrospectively, we could have taken $\tau = \tau(Q)$ in Lemma 12.4.)

We have shown that each of the maps θ_{γ_i} or ψ_{Y_i} restricted to the i-th face of hull(Q) is injective. It follows that the map $\psi^*_\tau : \text{hull}(Q) \to T^*(\tau)$, and hence $\omega^*_\tau : \text{hull}(Q) \to T^*(\tau)$, is injective.

Up until now, we have just assumed (A1)–(A10). To proceed, we will need also to assume the distance formula (B1) given in Section 7. As a consequence, we can weaken the hypotheses of Lemma 10.2 as follows:

Lemma 12.5. *If $a, b \in \mathcal{M}(\Sigma)$ are weakly (X, r)-related, then $\rho(a, b)$ agrees with $\rho_X(a, b)$ up to linear bounds depending only on r and on $\xi(\Sigma)$.*

Proof. The only contribution to the distance formula (property (B1)) in $\mathcal{M}(\Sigma)$ comes from subsurfaces of X, and so gives the same answer in $\mathcal{M}(X)$ up to linear bounds. $\square$

Lemma 12.6. *Let $Q \subseteq \mathcal{M}^*(\Sigma)$ be a ξ-cube. Then $Q \subseteq T^*(\tau(Q))$.*

Proof. Suppose, for contradiction, that $a \in Q \setminus T^*(\tau)$. We can write hull$(Q)$ as a direct product of the intervals $[a, a_i]$, where $\{a, a_i\}$ is the i-th side of Q containing a. By Lemma 2.1, $\mathcal{M}^*(\Sigma)$ is locally convex (since every component is), so there is a convex neighbourhood C of a with $C \cap T^*(\tau) = \varnothing$. Since $\mathcal{M}^*(\Sigma)$ has no isolated points, we can find some $b_i \in ([a, a_i] \cap C) \setminus \{a\}$. Now, since ω^*_τ is injective, $\omega^*_\tau b_i \neq \omega^*_\tau a$. Let W_i be a wall of $\mathcal{M}^*(\Sigma)$ separating $\omega^*_\tau b_i$ from $\omega^*_\tau a$. Since ω^*_τ is a gate map to $T^*(\tau)$, this also separates b_i from a. Also, since C and $T^*(\tau)$ are convex, there is another wall W separating C from $T^*(\tau)$. (Any two disjoint convex subsets of a median algebra are separated by a wall.) We see that the walls, $W, W_1, \dots, W_\xi$, all pairwise cross. We deduce that $\mathcal{M}^*(\Sigma)$ has rank at least $\xi + 1$, contradicting Lemma 6.6. $\square$

Lemma 12.7. *Suppose that $\gamma \in D^0_i(Q)$ and that $X \in D_j(Q)$ is a complexity-1 subsurface. If $i \neq j$, then either $\gamma \wedge X$, or there is some (unique) $\beta \in D^0_j(Q)$ with $\beta \prec X$.*

Proof. If γ is not disjoint from X, then we must have $\gamma \prec X$ or $\gamma \pitchfork X$, but the first possibility is ruled out by Lemma 12.3. Since X has complexity 1, either $\gamma \pitchfork Y$ for all $Y \preceq X$, or else there is a unique $\beta \prec X$ such that $\gamma \pitchfork Y$ whenever $Y \prec X$ and $Y \neq X(\beta)$. (In the former case, $\gamma \cap X$ cuts X into a collection of discs, and in the later it cuts X into discs together with one annulus, and we take β to be its core curve.) We claim that the latter case holds, and that $\beta \in D_j(Q)$.

Let ϕ_ζ, c, d, c, d' be as in the proof of Lemma 12.2. In the first case above, we follow the argument of Lemma 12.3 to derive a contradiction. In the second case, let $\beta_\zeta \to \beta$. For almost all ζ, we have $\beta_\zeta \prec X_\zeta$ and $\gamma_\zeta \pitchfork X_\zeta$. Given such ζ, suppose that $Y_\zeta \preceq X_\zeta$ and $Y \neq X(\beta_\zeta)$. Then almost always, $\sigma_\zeta(\phi_\zeta c', \phi_\zeta d')$ is bounded. (Otherwise, since $\gamma_\zeta \pitchfork Y_\zeta$ we derive a contradiction, as in the proof of Lemma 12.2.) Given any $e, f \in [\phi c, \phi d]$, we see that $\sigma_{Y_\zeta}(e_\zeta, f_\zeta)$ is (almost always) bounded. Thus, intrinsically to $\mathcal{M}(X_\zeta)$, we see that $\psi_{X_\zeta} e_\zeta$ and $\psi_{X_\zeta} f_\zeta$ are weakly β_ζ-related. It follows that $\sigma_{\beta_\zeta}(e_\zeta, f_\zeta) \asymp \rho_{X_\zeta}(e_\zeta, f_\zeta)$. Thus, if $e \neq f$, then since $X \in D_j(Q)$, we have $\rho_X(e, f) \neq 0$, so $\sigma_\beta(e, f) \neq 0$. It follows that $\theta^*_\beta|[\phi c', \phi d']$ is injective. In other words, $\beta \in D^0_j(Q)$ as claimed. □

We can summarise what we have shown as follows. Recall that $D_i(Q)$ is the set of subsurfaces, X, for which $\psi^*_X|[c, d]$ is injective for some (or equivalently any) i-th side, $[c, d]$, of Q.

Proposition 12.8. *For any i, the set $D^0_i(Q)$ is either empty or consists of a single curve $\gamma_i \in \mathcal{U}\mathbb{G}^0$. If it is empty, then there is a unique complexity-1 subsurface $Y_i \in D_i(Q)$. The set of γ_i are all disjoint, and they form a big multicurve $\tau(Q)$. The Y_i are also disjoint, and are precisely the complexity-1 components of $\tau(Q)$.*

We note, in particular, that γ_i or Y_i is completely determined intrinsically by any i-th side of Q, without reference to Q itself.

13. Flats in $\mathbb{M}^*(\Sigma)$

In this section, we restrict to the case where $\mathcal{M}(\Sigma) = \mathbb{M}(\Sigma)$ is the marking graph, and consider flats in the (extended) asymptotic cone. The parameters now depend only on $\xi(\Sigma)$.

First, we consider a particular case arising from complete multicurves. Suppose that $\tau \subseteq \mathcal{U}\mathbb{G}^0$ is a (nonstandard) complete multicurve. In other words, τ has ξ components and cuts Σ into $S_{0,3}$'s. In this case, each factor is a copy of $\mathbb{R}^*$, so $T^*(\tau)$ is isomorphic to $(\mathbb{R}^*)^\xi$. We refer to $T^*(\tau)$ as an *extended Dehn twist flat.*

More generally, if τ is big (that is each component of the complement is an $S_{0,3}$, $S_{0,4}$, or $S_{1,1}$), then again $\mathcal{U}\mathcal{X}(\tau)$ has ξ elements, and $T^*(\tau)$ is a direct product of ξ $\mathbb{R}^*$-trees.

If $X \in \mathcal{U}\mathcal{X}$, then $\mathbb{M}^*(X)$ and $\mathbb{G}^*(X)$ have preferred basepoints. These are defined as follows. Fix any $a \in \mathbb{M}(\Sigma)$ and let $e_X \in \mathbb{M}^*(X)$ be the limit of the points

$\psi_{X_\zeta}(a) \in \mathbb{M}^*(X_\zeta)$. This limit is independent of a. We similarly define $f_X \in \mathbb{G}^*(\Sigma)$ as the limit of $\theta_{X_\zeta}(a)$ (or equivalently, as $f_X = \chi^* e_X$). Let $\mathbb{M}^\infty(X)$ and $\mathbb{G}^\infty(X)$ be the components containing e_X and f_X respectively. Using Lemma 11.1, one sees that these are isomorphic to the asymptotic cones defined intrinsically on a standard surface of the topological type of X (unless X is an annulus, in which case, they are both isometric copies of $\mathbb{R}$). Note that $\theta^*_X(\mathbb{M}^\infty(\Sigma)) \subseteq \mathbb{G}^\infty(\Sigma)$. We will denote the restriction of θ^*_X to $\mathbb{M}^\infty(\Sigma)$ by θ^∞_X.

If τ is a complete multicurve, we write $T^\infty(\tau) = T^*(\tau) \cap \mathbb{M}^\infty(\Sigma)$. This is either empty or isomorphic to $\mathbb{R}^\xi$. In the latter case, this is naturally identified with $\mathcal{T}^\infty(\tau)$ — the direct product of $\mathbb{M}^\infty(X)$ for $X \in \mathcal{U}\mathcal{X}_T(\tau)$.

Definition. A *Dehn twist flat* in $\mathbb{M}^\infty(\Sigma)$ is a nonempty set of the form $T^*(\tau) \cap \mathbb{M}^\infty(\Sigma)$, where $\tau \subseteq \mathcal{U}\mathbb{G}^0(\tau)$ is a complete multicurve.

(We will explain the general term "flat" in this context below.)

By Lemma 11.1, up to the action of $\mathcal{U}\,\mathrm{Map}(\Sigma)$, we can take τ to be standard. One way to construct $T^\infty(\tau)$ in this case is as follows. Recall that $G(\tau) \cong \mathbb{Z}^\xi$ is the subgroup of $\mathrm{Map}(\Sigma)$ generated by Dehn twists about the components of τ. Let a be any element of $\mathbb{M}(\Sigma)$. The orbit, Ga, is a bounded Hausdorff distance from $T(\tau)$, and so $T^\infty(\tau)$ is the limit of Ga in the asymptotic cone $\mathbb{M}^\infty(\Sigma)$. The natural map from $\mathbb{Z}^\xi \cong G$ to Ga limits on an isomorphism from $\mathbb{R}^\xi$ to $T^\infty(\tau)$, where we view $\mathbb{R}^\xi$ as the asymptotic cone of $\mathbb{Z}^\xi$.

More generally, if τ is a big multicurve, then $T^\infty(\tau) = T^*(\tau) \cap \mathbb{M}^\infty(\Sigma)$ is either empty or a direct product of ξ $\mathbb{R}$-trees. In the latter case, it will contain many flats. We aim to show that every (maximal dimensional) flat in $\mathbb{M}^\infty(\Sigma)$ has this form. First, we consider the case of an $S_{1,1}$ or $S_{0,4}$.

Suppose that Σ is an $S_{1,1}$ or $S_{0,4}$. In this case, $\mathbb{G}(\Sigma)$ is a Farey graph, and (up to quasi-isometry) $\mathbb{M}(\Sigma)$ is the dual 3-valent tree (that is the dual to the Farey complex obtained by attaching a 2-simplex to every 3-cycle in $\mathbb{G}(\Sigma)$). To each $\gamma \in \mathbb{G}^0(\Sigma)$ we can associate a bi-infinite geodesic, or *axis*, in $\mathbb{M}(\Sigma)$. Up to bounded Hausdorff distance, we can identify this axis with the space $T(\gamma) = T(\{\gamma\})$ defined in Section 9, which we can, in turn, identify up to quasi-isometry, with $\mathbb{M}(\gamma) = \mathbb{G}(\gamma)$. Any two distinct axes meet in at most a single edge of $\mathbb{M}(\Sigma)$.

As noted before, in this case, $\mathbb{M}^*(\Sigma)$ and $\mathbb{G}^*(\Sigma)$ are both $\mathbb{R}^*$-trees. If $\gamma \in \mathcal{U}\mathbb{G}^0(\Sigma)$, we get a closed convex subset $T^*(\gamma) \subseteq \mathbb{M}^*(\Sigma)$, which can be identified with $\mathbb{M}^*(\gamma) = \mathbb{G}^*(\gamma) \cong \mathbb{R}^*$. If $\alpha, \beta \in \mathcal{U}\mathbb{G}^0(\Sigma)$ are distinct, then $T^*(\alpha) \cap T^*(\beta)$ consists of at most one point. The gate map $\omega^*_\gamma : \mathbb{M}^*(\Sigma) \to T^*(\gamma)$ is the limit of subsurface projection.

We now want to describe flats more generally. In this context, we make the following definition:

Definition. A *flat* in $\mathbb{M}^*(\Sigma)$ is a closed convex subset median isomorphic to $\mathbb{R}^\xi$.

Note that, in the case of a median metric space, this notion is equivalent to the notion of a flat as defined in Section 3. (In particular, "flats" are always assumed to have maximal rank.) In fact, we know that $\mathbb{M}^*(\Sigma)$ is bi-Lipschitz equivalent to a median metric space, so with respect to this median metric, the two notions coincide.

Let $\Phi \subseteq \mathbb{M}^\infty(\Sigma)$ be a flat. We identify Φ with $\mathbb{R}^\xi$ via a median isomorphism. Given $i \in \{1, \dots, \xi\}$, let $L_i \subseteq \Phi$ be an i-th coordinate line. (Note that two such are parallel. Moreover, they are determined up to permutation of the indices i.) Let $D_i(\Phi) = D(L_i)$, that is, the set of $X \in \mathcal{UX}(\Sigma)$ such that $\psi_X^\infty | L_i$ is injective. (This is independent of the choice of L_i.) We similarly define $D_i^0(\Phi) \subseteq \mathcal{U}\mathbb{G}^0(\Sigma)$, which we can identify as a subset of $D_i(\Phi)$.

We now bring Proposition 12.8 into play. Note that if Q is any ξ-cube in Φ, then $D_i(Q) \supseteq D_i(\Phi)$. In fact, there is some ξ-cube $Q_0 \subseteq \Phi$, with $D_i(Q_0) = D_i(\Phi)$ for all i, and so $D_i(Q) = D_i(\Phi)$ for any cube in Φ bigger than Q_0 (that is, with $Q_0 \subseteq \mathrm{hull}(Q)$). In particular, $|D_i^0(\Phi)| \le 1$. Let $I(\Phi) = \{i \mid D_i^0(\Phi) \neq \varnothing\}$. If $i \in I(\Phi)$, write $D_i^0(\Phi) = \{\gamma_i\}$, and let $\tau(\Phi) = \{\gamma_i \mid i \in I(\Phi)\}$. Thus, $\tau = \tau(\Phi) = \tau(Q_0)$ is a big multicurve. If Q is any bigger cube, then Lemma 12.6 tells us that $Q \subseteq T^\infty(\tau)$. Since $\mathrm{hull}(Q)$ is exhausted by such hulls, we conclude:

Proposition 13.1. *If $\Phi \subseteq \mathbb{M}^\infty(\Sigma)$ is a flat, then $\tau(\Phi)$ is a big multicurve, and $\Phi \subseteq T^\infty(\tau(\Phi))$. Moreover, if $Y \in \mathcal{UX}_N(\tau(\Phi))$, then $Y \in D_i(\Phi)$ for some $i \in \{1, \dots, \xi\} \setminus I(\Phi)$.*

Note also, as in Proposition 12.8, that each $Y \in \mathcal{UX}(\tau)$ lies in $D_i(\Phi)$ for some unique $i \notin I(\Phi)$.

Note that, applying Lemma 12.7 to a large cube in Φ, we see that if $\gamma \in D_i^0(\Phi)$, then for all $j \in \{1, \dots, \xi\} \setminus I(\Phi)$ if $X \in D_j(\Phi)$ is an $S_{1,1}$ or $S_{0,4}$, and $i \neq j$, then $\gamma \wedge X$ (since the second possibility is ruled out by the fact that $D_j^0(\Phi) = \varnothing$).

Next, we aim to describe when two flats meet in a codimension-1 plane (necessarily a coordinate subspace).

Lemma 13.2. *Let Φ_0, Φ_1, be two flats with $\Phi_0 \cap \Phi_1$ a codimension-1 coordinate plane. Then $\tau = \tau(\Phi_0) \cap \tau(\Phi_1)$ is a big multicurve. Moreover, $|\tau(\Phi_i) \setminus \tau| \le 1$. If $\beta_0 \in \tau(\Phi_0) \setminus \tau$ and $\beta_1 \in \tau(\Phi_1) \setminus \tau$ then $\beta_0 \neq \beta_1$ and β_0 and β_1 lie in the same complementary component of τ.*

Proof. Choose coordinates on Φ_0 and Φ_1 so that $\Phi_0 \cap \Phi_1$ is a plane orthogonal to the 1st axis, and so that the other coordinates agree on $\Phi_0 \cap \Phi_1$. Write $I_i = I(\Phi_i)$ and $\tau_i = \tau(\Phi_i)$. Let $\tau = \tau_0 \cap \tau_1$. Now $I_0 \setminus \{1\} = I_1 \setminus \{1\}$ (since these sets are determined by lines in $\Phi_0 \cap \Phi_1$). The only case we need to consider is where $1 \in I_0 \cap I_1$ (otherwise, at least one of τ_0 or τ_1 agrees with τ and the statement follows). We aim to show that τ_0 and τ_1 differ only inside a complexity-1 component of $\Sigma \setminus \tau$, and it will follow that τ is big.

So suppose that $1 \in I_0 \cap I_1$. Then $\tau_0 = \tau \cup \{\beta_0\}$ and $\tau_1 = \tau \cup \{\beta_1\}$. Let $Y_i \in \mathcal{UX}_N(\tau)$ be the component containing β_i.

If $Y_0 \neq Y_1$, then $Y_0 \in \mathcal{UX}_N(\tau_1)$, so $Y_0 \in D_i(\Phi_1)$ for some $i \neq 1$ (as observed after Proposition 13.1). But $D_i(\Phi_0) = D_i(\Phi_1)$. In other words, we have $\beta_0 \prec Y_0$, $\beta_0 \in D_1(\Phi_0)$, $Y_0 \in D_i(\Phi_0)$ and Y_0 is an $S_{1,1}$ or $S_{0,4}$. It now follows that $\beta_0 \wedge Y_0$, giving a contradiction.

Thus, $Y_0 = Y_1 = Y$, say. Since $\Phi_0 \neq \Phi_1$, we must have $\beta_0 \neq \beta_1$. We claim that Y is an $S_{1,1}$ or an $S_{0,4}$. For suppose not. We use the fact that τ_0 and τ_1 are big. Either $\beta_0 \pitchfork \beta_1$ or $\beta_0 \wedge \beta_1$. In the former case, we have $\beta_0 \pitchfork Z$ for some $Z \in \mathcal{UX}(\tau_1)$ and we get a contradiction as before. In the latter case, we have $\beta_0 \prec W$ for some $W \in \mathcal{UX}(\tau_1)$ and we derive a similar contradiction.

Thus, Y is an $S_{1,1}$ or $S_{0,4}$. Since τ_0 and τ_1 are big, and differ only in the curves β_0, β_1, it follows that τ is big. □

Elaborating on the above proof, we see that there are essentially three possibilities (up to swapping Φ_0 and Φ_1). Let us suppose that Φ_0 and Φ_1 differ in the first coordinate. We have one of the following:

(1) $\tau(\Phi_0) = \tau(\Phi_1) = \tau$. In this case, there is some $Y \in \mathcal{UX}_N(\tau)$ corresponding to the first factor of both $T(\tau_0)$ and $T(\tau_1)$, so that Φ_0 and Φ_1 project to lines meeting in a single point in the $\mathbb{R}$-tree $\mathbb{M}^\infty(Y)$.

(2) $\tau(\Phi_0) = \tau$ and $\tau(\Phi_1) = \tau \cup \{\beta\}$. Let $Y \in \mathcal{UX}_N(\tau)$ be the component containing β. In the $\mathbb{R}$-tree $\mathbb{M}^\infty(Y)$, Φ_1 projects to the axis corresponding to β, and Φ_0 projects to a line meeting this axis in a single point.

(3) $\tau(\Phi_0) = \tau \cup \{\beta_0\}$ and $\tau(\Phi_1) = \tau \cup \{\beta_1\}$. Let $Y \in \mathcal{UX}(\tau)$ be the component containing β_0 and β_1. Then Φ_0 and Φ_1 respectively project to the axes in $\mathbb{M}^\infty(Y)$ corresponding to β_0 and β_1. These axes intersect in a single point.

We next want to characterise Dehn twist flats.

Lemma 13.3. *Suppose that $\Phi \subseteq \mathbb{M}^\infty(\Sigma)$ is a flat. Suppose that for each i there is another flat $\Phi_i \subseteq \mathbb{M}^\infty(\Sigma)$ with $\Phi \cap \Phi_i$ a codimension-1 coordinate plane orthogonal to the i-th axis. Then Φ is a Dehn twist flat.*

In fact, it is enough to assume the hypothesis for those $i \in I(\Phi)$.

Proof. Suppose $i \in I(\Phi)$. Let $\gamma_i \in \tau(\Phi)$ be the corresponding curve. By Lemma 13.2 and subsequent discussion, we see that $\tau(\Phi_i)$ is obtained from $\tau(\Phi)$ by deleting γ_i and possibly replacing it by another curve in the complementary component of $\tau(\Phi) \setminus \{\gamma_i\}$ that contained γ_i. But $\tau(\Phi_i)$ is big, so either way, it follows that γ_i must lie in an $S_{1,1}$ or $S_{0,4}$ component of the complement of $\tau(\Phi) \setminus \{\gamma_i\}$. Put another way, γ_i bounds an $S_{0,3}$ component of $\Sigma \setminus \tau(\Phi)$ (possibly the same $S_{0,3}$) on each side. Since this holds for all $i \in I(\Phi)$ (that is for all components of $\tau(\Phi)$) it follows that each component of $\Sigma \setminus \tau(\Phi)$ is an $S_{0,3}$. In other words, $\tau(\Phi)$ is complete. □

For the converse, suppose that Φ is a Dehn twist flat. For simplicity, we can assume that $\tau = \tau(\Phi)$ is standard. Let $G = G(\tau) \subseteq \mathrm{Map}(\Sigma)$ be the subgroup generated by Dehn twists about the components of τ. Thus $G \cong \mathbb{Z}^\xi$. Let $\mathcal{U}G \leq \mathcal{U}\,\mathrm{Map}(\Sigma)$ be its ultraproduct, and let $\mathcal{U}^0 G = \mathcal{U}G \cap \mathcal{U}^0\,\mathrm{Map}(\Sigma)$. (Recall, from Section 5, that $\mathcal{U}^0\,\mathrm{Map}(\Sigma)$ is defined to be the setwise stabiliser of $\mathbb{M}^\infty(\Sigma)$.) Then $\mathcal{U}^0 G$ acts transitively on Φ, preserving the coordinate directions.

Lemma 13.4. *Suppose that Φ is a Dehn twist flat. Then if Θ is any codimension-1 coordinate subspace in Φ, then there is some Dehn twist flat Ψ with $\Theta = \Phi \cap \Psi$.*

Proof. For simplicity, we can assume $\tau = \tau(\Phi)$ to be standard. Let $\gamma \in \tau$ be the curve corresponding to the coordinate direction perpendicular to Θ. Let $Y \in \mathcal{U}\mathcal{X}(\tau \setminus \{\gamma\})$ be the component containing γ. Let $\gamma \in \mathbb{G}^0(\Sigma)$ be any other standard curve in Y. Now the axes of β and γ in $\mathbb{G}^\infty(Y)$ meet in a single point. Let $\tau' = (\tau \setminus \{\gamma\}) \cup \{\beta\}$, and let $\Psi = T(\tau')$. Then Ψ is a Dehn twist flat meeting Φ is a codimension-1 plane parallel to Θ. By the homogeneity of Φ described before the statement of the lemma, this is sufficient to prove the result. □

Putting the above together with Proposition 4.6, we get:

Proposition 13.5. *Suppose that $\Phi \subseteq \mathbb{M}^\infty(\Sigma)$ is a closed subset and that there is a homeomorphism $f : \mathbb{R}^\xi \to \Phi$ with the following property. For each codimension-1 coordinate plane $H \subseteq \mathbb{R}^\xi$ there is a closed subset $\Psi \subseteq \mathbb{M}^\infty(\Sigma)$ homeomorphic to $\mathbb{R}^\xi$ such that $f(H) = \Phi \cap \Psi$. Then Φ is a Dehn twist flat, and f is a median isomorphism. Moreover, every Dehn twist flat arises in this way.*

In particular, we see that the collection of Dehn twist flats is determined by the topology of $\mathbb{M}^\infty(\Sigma)$, as shown in [Behrstock, Kleiner, Minsky and Mosher 2012]. In fact, we only need an injective map. Moreover, we can take two different surfaces with the same complexity. In summary, we conclude:

Theorem 13.6. *Suppose that Σ and Σ' are compact surfaces with $\xi(\Sigma) = \xi(\Sigma') \geq 2$. Suppose that we have a continuous injective map $f : \mathbb{M}^\infty(\Sigma) \to \mathbb{M}^\infty(\Sigma')$ with closed image. If Φ is a Dehn twist flat in $\mathbb{M}^\infty(\Sigma)$, then $f(\Phi)$ is a Dehn twist flat in $\mathbb{M}^\infty(\Sigma')$.*

Note that this applies equally well to any components of $\mathbb{M}^*(\Sigma)$ and $\mathbb{M}^*(\Sigma')$, since they are all respectively isomorphic to $\mathbb{M}^\infty(\Sigma)$ and to $\mathbb{M}^\infty(\Sigma')$.

14. Controlling Hausdorff distance

We begin a general statement, which generalises a construction of [Behrstock, Kleiner, Minsky and Mosher 2012].

Let (M, ρ) be a metric space. Given subsets, $A, B, D \subseteq M$, we say that A, B are *r-close on D* if $A \cap D \subseteq N(B; r)$ and $B \cap D \subseteq N(A; r)$. (Thus r-close on M

means that the Hausdorff distance, $\mathrm{hd}(A, B)$, from A to B is at most r.) Let $\boldsymbol{t}$ be a positive infinitesimal, and let M^* be the extended asymptotic cone determined by $\boldsymbol{t}$. Given $e \in M^*$, let M_e^∞ be the component of M^* containing e. Let $\boldsymbol{r} = 1/\boldsymbol{t}$.

Let $\mathcal{UP}(M)$ be the ultrapower of the power set, $\mathcal{P}(M)$, of M. Given $\boldsymbol{A} \in \mathcal{UP}(M)$, let $\mathcal{U}\boldsymbol{A}$ and $A^* \subseteq M^*$ be the images of $\boldsymbol{A}$ under the natural maps $\mathcal{UP}(M) \to \mathcal{P}(\mathcal{U}(M)) \to \mathcal{P}(M^*)$ (as discussed in Section 5).

The following is a simple observation (a similar statement is used in [Behrstock, Kleiner, Minsky and Mosher 2012]).

Lemma 14.1. *Suppose that $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{UP}(M)$, and $\boldsymbol{e} \in \mathcal{U}\boldsymbol{A}$ (that is $e_\zeta \in A_\zeta$ for almost all ζ). Let $e \in M^*$ be the image of $\boldsymbol{e}$ in M^* (so that $e \in A^*$). Suppose that $\epsilon, R > 0$ are positive real numbers. Then A^*, B^* are ϵ-close on $N(e; R)$ if and only if, for all $R' > R$ and all $\epsilon' > \epsilon$, the sets A_ζ, B_ζ are $\epsilon' r_\zeta$-close on $N(e_\zeta; R' r_\zeta)$ for almost all ζ.*

In particular, if $A^* \cap M_e^\infty = B^* \cap M_e^\infty$, then for all $R > \epsilon > 0$, the sets A_ζ, B_ζ are almost always ϵr_ζ-close on $N(e_\zeta; R r_\zeta)$. (Here "almost" may depend on ϵ and R.) Note that, in the above, only the component, M_e^∞, of M^* containing e is relevant.

Lemma 14.2. *Suppose that for all $R > \epsilon > 0$ there is some $e \in A^*$ such that A^*, B^* are ϵ-close on $N(e; R)$. Then, there is some component M^0 of M^* such that $A^* \cap M^0 = B^* \cap M^0 \neq \varnothing$.*

Proof. Given any $n \in \mathbb{N}$, there is some e_n such that A^*, B^* are $1/(2n)$-close on $N(\boldsymbol{e}_n; 2n)$. Write $e_n = (e_{n,\zeta})_\zeta$. Let $\mathcal{Z}_n$ be the set of $\zeta \in \mathcal{Z}$ such that A_ζ, B_ζ are r_ζ/n-close on $N(e_{n,\zeta}; n r_\zeta)$. Thus, for all n, $\mathcal{Z}_n$ has measure 1. Given $\zeta \in \mathcal{Z}$, let $m(\zeta) = \max(\{n \mid \zeta \in \mathcal{Z}_n\} \cup \{0\}) \in \mathbb{N} \cup \{\infty\}$. Let $p : \mathcal{Z} \to \mathbb{N}$ be any map with $p(\zeta) \to \infty$ (for example, any injective map from $\mathcal{Z}$ to $\mathbb{N}$). Let $n(\zeta) = \min\{m(\zeta), p(\zeta)\} \in \mathbb{N}$. Note that $n(\zeta) \to \infty$ (since for any $n \in \mathbb{N}$, $p(\zeta) > n$ almost always, and $\zeta \in \mathcal{Z}_n$ so that $m(\zeta) > n$ almost always). Let $e_\zeta = e_{n(\zeta),\zeta}$, and let e be the image of $(e_\zeta)_\zeta$ in A^*. Now, for all n, A_ζ, B_ζ are almost always r_ζ/n-close on $N(e_\zeta; n r_\zeta)$, so A^*, B^* are $1/n$-close on $N(e; n)$. Since this holds for all n, we have $A^* \cap M_e^\infty = B^* \cap M_e^\infty \neq \varnothing$, as required. □

Suppose now that $\mathcal{E}$ and $\mathcal{F}$ are collections of subsets of M. We write $\mathcal{UE}$ and $\mathcal{UF}$ for their respective ultrapowers.

We suppose:

(S1) E is (coarsely) connected for all $E \in \mathcal{E}$.

(S2) If $\boldsymbol{F}, \boldsymbol{F}' \in \mathcal{UF}$ and there is some component M^0 of M^* such that $F^* \cap M^0 = (F')^* \cap M^0 \neq \varnothing$, then $\boldsymbol{F} = \boldsymbol{F}'$.

(S3) For all $\boldsymbol{E} \in \mathcal{UE}$, and for all components M^0 of M^*, there is some $\boldsymbol{F} \in \mathcal{UF}$ such that $E^* \cap M^0 = F^* \cap M^0$.

In fact, we only really require (S3) if $E^* \cap M^0 \neq \emptyset$.

(In (S1), "coarsely connected" can be taken to mean that $N(E; s)$ is connected for some fixed $s \in [0, \infty) \subseteq \mathbb{R}$.)

Lemma 14.3. *If $\mathcal{E}, \mathcal{F}$ satisfy (S1)–(S3) above, then there is some $k > 0$ such that for all $E \in \mathcal{E}$, there is some $F \in \mathcal{F}$, such that* $\mathrm{hd}(E, F) \leq k$.

Proof. Suppose not. Let $\epsilon > 0$. Given any $\zeta \in \mathcal{Z}$, there is some $E_\zeta \in \mathcal{E}$ such that for all $F \in \mathcal{F}$, $\mathrm{hd}(E_\zeta, F) > \epsilon r_\zeta$. Let $\boldsymbol{E} = (E_\zeta)_\zeta \in \mathcal{U}\mathcal{E}$. Let e_ζ be any element of E_ζ (so that $e \in E^*$). By (S3), there is some $\boldsymbol{F} \in \mathcal{U}\mathcal{F}$ such that $E^* \cap M^\infty_e = F^* \cap M^\infty_e$. In particular, for all $R > 4\epsilon$, we have that E_ζ, F_ζ are almost always $\epsilon r_\zeta/2$-close on $N(e_\zeta; 2Rr_\zeta)$. But $\mathrm{hd}(E_\zeta, F_\zeta) > \epsilon r_\zeta$, so there is some $e'_\zeta \in E_\zeta$ such that E_ζ, F_ζ are not $\epsilon r_\zeta/2$-close on $N(e'_\zeta; 2Rr_\zeta)$. By (S1), we can find $q_\zeta, q'_\zeta \in E_\zeta$ with $\rho(q_\zeta, q'_\zeta)$ bounded such that E_ζ, F_ζ are $\epsilon r_\zeta/2$-close on $N(q_\zeta; 2Rr_\zeta)$ but not on $N(q'_\zeta; 2Rr_\zeta)$. But by (S3) again, there is almost always some $F'_\zeta \in \mathcal{F}$ such that E_ζ, F'_ζ are $\epsilon r_\zeta/2$-close on $N(q'_\zeta; 2Rr_\zeta)$. Clearly $F'_\zeta \neq F_\zeta$. It follows that F_ζ, F'_ζ are ϵr_ζ-close on $N(q_\zeta; Rr_\zeta) \subseteq N(q_\zeta; 2Rr_\zeta) \cap N(q'_\zeta; 2Rr_\zeta)$. (Almost always, $\rho(q_\zeta, q'_\zeta) < Rr_\zeta$.) Let $\boldsymbol{F}' = (F'_\zeta)_\zeta$. We see that F^*, $(F')^*$ are ϵ-close on $N(q; R)$. Since $R > 4\epsilon > 0$ were arbitrary, it follows from Lemma 14.2 that there is some component, M^0, of M^* such that $F^* \cap M^0 = (F')^* \cap M^0 \neq \emptyset$. By (S2), we have $\boldsymbol{F} = \boldsymbol{F}'$. But $F'_\zeta \neq F_\zeta$ almost always, giving a contradiction. □

We have the following criterion to verify (S2).

Given $A, B \subseteq M$, we say that B *linearly diverges* from A if there are constants $k, t \geq 0$ such that for all $r \geq 0$ and all $x \in B$, there is some $y \in B$ with $\rho(y, A) \geq r$ and $\rho(x, y) \leq kr + t$. We say that a collection $\mathcal{F}$ of subsets of M *linearly diverges* if given any distinct $A, B \in \mathcal{F}$, B linearly diverges from A, with k, t uniform over $\mathcal{F}$.

Lemma 14.4. *If a family $\mathcal{F}$ of subsets linearly diverges, then it satisfies (S2) above.*

Proof. Suppose that $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{U}\mathcal{F}$ and $A^* \cap M^0 = B^* \cap M^0 \neq \emptyset$, for some component M^0 of M^*. If $e \in B^* \cap M^0$, then we have $e_\zeta \in B_\zeta$ with $e_\zeta \to e$. Setting $\epsilon = 1$ and $R > 3k$, we have that A_ζ and B_ζ are almost always r_ζ-close on $N(e; Rr_\zeta)$. If $A_\zeta \neq B_\zeta$, then there is some $y \in B_\zeta$, with $\rho(y, A_\zeta) \geq 2r_\zeta$ and $\rho(e_\zeta, y) \leq 2kr_\zeta + t < 3kr_\zeta$ almost always. Thus, $y \in N(e; Rr_\zeta)$, so we get the contradiction that $\rho(y, A_\zeta) \leq r_\zeta$. Thus $A_\zeta = B_\zeta$ almost always, that is, $\boldsymbol{A} = \boldsymbol{B}$. □

Finally, we apply this to the marking complexes to show that coarse Dehn twist flats get sent (close) to coarse Dehn twist flats under a quasi-isometric embedding.

Suppose that Σ and Σ' are compact surfaces with $\xi = \xi(\Sigma) = \xi(\Sigma')$. Suppose that $\phi : \mathbb{M}(\Sigma) \to \mathbb{M}(\Sigma')$. This gives rise to a continuous map $\phi^* : \mathbb{M}^*(\Sigma) \to \mathbb{M}^*(\Sigma')$ with closed image. In fact, each component $\mathbb{M}^*_e(\Sigma)$ of $\mathbb{M}^*(\Sigma)$ gets sent into the component $\mathbb{M}^*_{\phi^*(e)}(\Sigma')$ of $\mathbb{M}^*(\Sigma')$. Moreover, distinct components get sent into distinct components.

Let $\mathcal{F}(\Sigma)$ be the set of coarse twist flats, $T(\tau)$, as τ ranges over all complete multicurves, τ. This satisfies (S1). Also, it is linearly divergent, by Lemma 9.9, and so therefore satisfies (S2) by Lemma 14.4. Note that a Dehn twist flat in a component M^0 of $\mathbb{M}^*(\Sigma)$ is by definition a nonempty set of the form $F^* \cap M^0$ for some $\boldsymbol{F} \in \mathcal{U}\mathcal{F}(\Sigma)$. The same discussion applies to $\mathcal{F}(\Sigma')$.

Let $\mathcal{E} = \{\phi(F) \mid F \in \mathcal{F}(\Sigma)\}$. We claim that $\mathcal{E}, \mathcal{F}(\Sigma')$ satisfies (S3) with $M = \mathbb{M}(\Sigma')$.

Suppose $\boldsymbol{E} \in \mathcal{U}\mathcal{E}$. Then $\boldsymbol{E} = (\phi W_\zeta)_\zeta$, where $W_\zeta \in \mathcal{F}(\Sigma)$. Thus $E^* = \phi^* W^*$, where $\boldsymbol{W} = (W_\zeta)_\zeta$. Suppose that M^0 is a component of $\mathbb{M}^*(\Sigma')$ with $E^* \cap M^0 \neq \varnothing$. Choose any $e \in W^*$ with $\phi^* e \in M^0$. Thus, $M^0 = \mathbb{M}^*_{\phi^*(e)}(\Sigma')$. We see that $\phi^*(\mathbb{M}^*_e(\Sigma)) = M^0 \cap \phi^*(\mathbb{M}^*(\Sigma))$. Now $W^* \cap \mathbb{M}^*_e(\Sigma)$ is a Dehn twist flat in M^0, so by Theorem 13.6, $S^* \cap M^0 = \phi(W^*) \cap M^0 = \phi(W^*) \cap \mathbb{M}^*_e(\Sigma)$ is a Dehn twist flat in M^0. In other words, there is some $F \in \mathcal{U}\mathcal{F}(\Sigma')$ with $F^* \cap M^0 = E^* \cap M^0$. This verifies property (S3) for $\mathcal{E}, \mathcal{F}(\Sigma')$.

Lemma 14.5. *Suppose that Σ and Σ' are compact orientable surfaces with $\xi(\Sigma) = \xi(\Sigma') \geq 2$, and that $\phi : \mathbb{M}(\Sigma) \to \mathbb{M}(\Sigma')$ is a quasi-isometric embedding. Then there is some $k \geq 0$ such that if τ is a complete multicurve in Σ, then there is a complete multicurve τ' in Σ' such that* $\operatorname{hd}(T(\tau'), \phi T(\tau)) \leq k$.

Proof. We apply Lemma 14.3 to the sets $\mathcal{E} = \{\phi(F) \mid F \in \mathcal{F}(\Sigma)\}$ and $\mathcal{F} = \mathcal{F}(\Sigma')$. We have verified that $\mathcal{E}$ and $\mathcal{F}$ satisfy (S1)–(S3). □

As we have stated it (to keep the logic of the argument simpler) the bound k might depend on the particular map ϕ. In fact, it can be seen to depend only on ξ and the parameters of ϕ. For this, fix some parameters of quasi-isometry, and now take $\mathcal{E}$ to the set of all images $\phi(F)$, both as F ranges of the set of coarse Dehn twist flats, $\mathcal{F}(\Sigma)$, and as ϕ ranges over all quasi-isometric embeddings from $\mathbb{M}(\Sigma)$ to $\mathbb{M}(\Sigma')$ with these parameters. To verify (S3) we take $\boldsymbol{E} = (\phi_\zeta W_\zeta)_\zeta$ and apply Theorem 13.6, to the limiting map ϕ^* of $(\phi_\zeta)_\zeta$. The same argument now gives us a uniform constant k independent of any particular ϕ. (See the remark at the end of Section 6.)

15. Rigidity of the marking graph

In this section, we show that, modulo a few exceptional cases, a quasi-isometric embedding between mapping class groups is a bounded distance from a left multiplication (hence a quasi-isometry). This strengthens the result of [Hamenstädt 2005; Behrstock, Kleiner, Minsky and Mosher 2012].

Let (X, ρ) be a geodesic space. Given $A, B \subseteq X$, write $A \sim B$ to mean that $\operatorname{hd}(A, B) < \infty$. Clearly, this is an equivalence relation, and we write $\mathcal{B}(X)$ for the set of $\sim$-classes. Let $\mathcal{Q}(X) \subseteq \mathcal{B}(X)$ denote the set of $\sim$-classes of images of bi-infinite quasigeodesics.

If $A, B \in \mathcal{B}(X)$, we write $A \leq B$ to mean that some representative of A is contained in some representative of B. This "coarse inclusion" defines a partial order on $\mathcal{B}(X)$.

We say that two sets $A, B \subseteq X$ have *coarse intersection* if there is some $r \geq 0$ such that for all $s \geq r$, $N(A; r) \cap N(B; r) \sim N(A; s) \cap N(B; s)$ (cf., [Behrstock, Kleiner, Minsky and Mosher 2012]). Clearly, this depends only on the $\sim$-classes of A and B, and determines an element of $\mathcal{B}(X)$, denoted $A \wedge B$.

Note that if $\phi : X \to Y$ is a quasi-isometric embedding of X into another geodesic space, Y, then ϕ induces an injective map from $\mathcal{B}(X)$ to $\mathcal{B}(Y)$. Note that this respects inclusion and coarse intersection.

Suppose now that Γ is a group acting by isometry on X. We say that Γ acts *discretely* if for some (or equivalently any) $a \in X$ and any $r \geq 0$, the set $\{g \in \Gamma \mid \rho(a, ga) \leq r\}$ is finite. (In other words, a has finite stabiliser and locally finite orbit.) We will assume the action to be discrete here.

Any subgroup, $G \leq \Gamma$ determines an element $B(G)$ of $\mathcal{B}(X)$, namely the $\sim$-class of any G-orbit. If $G \leq H \leq \Gamma$, then $B(G) \leq B(H)$, with equality if and only if G has finite index in H. In fact, if $G, H \leq \Gamma$, then $B(G) = B(H)$ if and only if G, H are commensurable in Γ (i.e., $G \cap H$ has finite index in both G and H). More generally, for any $G, H \leq \Gamma$, $B(G)$ and $B(H)$ have coarse intersection, and $B(G \cap H) = B(G) \cap B(H)$. Note that $B(G)$ is the class of bounded sets if and only if G is finite. Also, the class $B(G)$ contains a bi-infinite quasigeodesic if and only if G is two-ended (virtually $\mathbb{Z}$) and undistorted in X.

Now, let Σ be a compact surface. Note that $\mathrm{Map}(\Sigma)$ acts discretely on $\mathbb{M}(\Sigma)$. If $\tau \subseteq \Sigma$ is a multicurve, let $G(\tau) \subseteq \mathrm{Map}(\Sigma)$ be the group generated by twists about the elements of τ. Thus, $G(\tau) \cong \mathbb{Z}^{|\tau|}$. Write $B(\tau) = B(G(\tau))$. Note that $B(\tau)$ determines τ uniquely. If τ, τ' are multicurves, then $G(\tau \cap \tau') = G(\tau) \cap G(\tau')$, and so $B(\tau \cap \tau') = B(\tau) \wedge B(\tau')$. Note that if τ is a complete multicurve, then $B(\tau)$ is the class of the coarse Dehn twist flat, $T(\tau)$.

Now if $\gamma \in \mathbb{G}^0(\Sigma)$, then we can always find complete multicurves τ, τ' with $\tau \cap \tau' = \{\gamma\}$. (In fact, we can choose τ, τ' with $\iota(\tau, \tau')$ uniformly bounded.) If $\gamma, \delta \in \mathbb{G}^0(\Sigma)$, then γ, δ are equal or adjacent in $\mathbb{G}(\Sigma)$ if and only if there is a complete multicurve, τ containing both γ and δ. Thus, $B(\gamma), B(\delta) \leq B(\tau)$.

Suppose now that Σ, Σ' are compact surfaces with $\xi(\Sigma) = \xi(\Sigma') \geq 2$. Suppose that $\phi : \mathbb{M}(\Sigma) \to \mathbb{M}(\Sigma')$ is a quasi-isometric embedding.

Suppose that $\tau \subseteq \Sigma$ is a complete multicurve. Now Lemma 14.5 gives us a complete multicurve $\tau' \subseteq \Sigma'$ with $\mathrm{hd}(T(\tau'), \phi T(\tau))$ bounded and, in particular, finite. Thus, $\phi(B(\tau)) = B(\tau')$. Moreover, this determines τ' uniquely, and we denote it by $\theta\tau$. Note that, from the remark following Lemma 14.5, we see that the bound depends only on the complexity of the surfaces and the parameters of quasi-isometry.

Suppose that $\gamma \in \mathbb{G}^0(\Sigma)$. Choose complete multicurves τ, τ' with $\tau \cap \tau' = \{\gamma\}$. Thus $B(\tau) \wedge B(\tau') = B(\gamma) \in \mathcal{Q}(\mathbb{M}(\Sigma))$, and so $B(\theta\tau) \wedge B(\theta\tau') \in \mathcal{Q}(\mathbb{M}(\Sigma'))$. It follows that $\theta\tau \cap \theta\tau'$ consists of a single curve, $\delta \in \mathbb{G}^0(\Sigma')$. Note that $B(\delta) = \phi(B(\gamma))$, and we see that δ is determined by γ. We write it as $\theta\gamma$. We have shown that there is a unique map, $\theta : \mathbb{G}^0(\Sigma) \to \mathbb{G}^0(\Sigma')$ such that $B(\theta\gamma) = \phi B(\gamma)$ for all $\gamma \in \mathbb{G}^0(\Sigma)$. Since $\phi : \mathcal{B}(\mathbb{M}(\Sigma)) \to \mathcal{B}(\mathbb{M}(\Sigma'))$ is injective, it follows that θ is injective.

Moreover, if γ, δ are equal or adjacent in $\mathbb{G}(\Sigma)$, then $\gamma, \delta \in \tau$ for some complete multicurve τ. So $B(\gamma), B(\delta) \leq B(\tau)$, so $B(\theta\gamma), B(\theta\delta) \leq B(\theta\tau)$, and so $\theta\gamma, \theta\delta$ are equal or adjacent in $\mathbb{G}(\Sigma')$. In other words, θ gives an injective embedding of $\mathbb{G}(\Sigma)$ into $\mathbb{G}(\Sigma')$.

We now use the following fact from [Shackleton 2007]:

Theorem 15.1. *Suppose that Σ and Σ' are compact surfaces with $\xi(\Sigma) = \xi(\Sigma') \geq 4$. If $\theta : \mathbb{G}(\Sigma) \to \mathbb{G}(\Sigma')$ is an injective embedding, then $\Sigma = \Sigma'$ and there is some $g \in \mathrm{Map}(\Sigma)$ such that $\theta\gamma = g\gamma$ for all $\gamma \in \mathbb{G}^0(\Sigma)$. The same conclusion holds if Σ, Σ' are both an $S_{2,0}$; if both an $S_{0,6}$; if both an $S_{0,5}$; or if at least one is an $S_{1,3}$, and the other has complexity $\xi = 3$.*

Applying this to our situation, we see that $\Sigma = \Sigma'$, and that there is some $g \in \mathrm{Map}(\Sigma)$ with $\theta\gamma = g\gamma$ for all $g \in \mathbb{G}^0(\Sigma)$. After postcomposing with g^{-1}, we may as well assume that g is the identity. In particular, it follows that $B(\tau) = \phi(B(\tau))$ for all complete multicurves τ in Σ. Now Lemma 14.5 gives us a uniform k such that $\mathrm{hd}(T(\tau'), \phi T(\tau)) \leq k$ for some multicurve τ' in Σ. But we now know that $\tau' = \tau$, and so we deduce that $\mathrm{hd}(T(\tau), \phi T(\tau)) \leq k$ for all multicurves, τ.

Now if $x \in \mathbb{M}(\Sigma)$, we can always find τ, τ' with $\tau \cap \tau' = \varnothing$, and with $\iota(\tau, \tau')$, $\rho(x, T(\tau))$ and $\rho(x, T(\tau'))$ all uniformly bounded. It follows that ϕx is a bounded distance from both $\phi T(\tau)$ and $\phi T(\tau')$ and so $\rho(\phi x, T(\tau))$ and $\rho(\phi x, T(\tau'))$ are also uniformly bounded. But $T(\tau)$ and $T(\tau')$ coarsely intersect in the class of bounded sets. Since there are only finitely many possibilities for the pair τ, τ' up to the action of $\mathrm{Map}(\Sigma)$ we can take the various constants to be uniform. This shows that $\rho(x, \phi x)$ is bounded.

We have shown:

Theorem 15.2. *Suppose that Σ and Σ' are compact surfaces with $\xi(\Sigma) = \xi(\Sigma') \geq 4$, and that $\phi : \mathbb{M}(\Sigma) \to \mathbb{M}(\Sigma')$ is a quasi-isometric embedding. Then $\Sigma = \Sigma'$ and there is some $g \in \mathrm{Map}(\Sigma)$ such that for all $a \in \mathbb{M}(\Sigma)$, we have $\rho(\phi a, ga) \leq k$, where k depends only on $\xi(\Sigma) = \xi(\Sigma')$ and the parameters of quasi-isometry of ϕ.*

(Note that if Σ, Σ' are compact surfaces and there is a quasi-isometric embedding of $\mathbb{M}(\Sigma)$ into $\mathbb{M}(\Sigma')$, then certainly $\xi(\Sigma) \leq \xi(\Sigma')$, since the complexity, $\xi = \xi(\Sigma)$, is the maximal dimension of a quasi-isometrically embedded copy of $\mathbb{R}^\xi$ in $\mathbb{M}(\Sigma)$. It is not clear when a quasi-isometric embedding exists if $\xi(\Sigma) < \xi(\Sigma')$.)

One can also describe the lower complexity cases. Note that complexity $\xi = 3$ corresponds to one of $S_{2,0}$, $S_{1,3}$ and $S_{0,6}$. Suppose that $\xi(\Sigma) = \xi(\Sigma') = 3$. Then the result of [Shackleton 2007], quoted as Theorem 15.1 here, tells us if $S_{1,3} \in \{\Sigma, \Sigma'\}$, then again $\Sigma = \Sigma'$ in which case, the conclusion of the theorem holds. Otherwise, it is necessary to assume that $\Sigma = \Sigma'$, and then the conclusion holds. Note that, in fact, the centre of $\mathrm{Map}(S_{2,0})$ is $\mathbb{Z}_2$, generated by the hyperelliptic involution. The quotient $\mathrm{Map}(S_{2,0})/\mathbb{Z}_2$ is isomorphic to $\mathrm{Map}(S_{0,6})$. Thus, $\mathbb{M}(S_{2,0})$ and $\mathbb{M}(S_{0,6})$ are quasi-isometric. Of course, the above allows us to describe the quasi-isometric embeddings between them up to bounded distance, as compositions of maps of the above type.

Suppose that $\xi(\Sigma) = \xi(\Sigma) = 2$. In this case $\Sigma \in \{S_{1,2}, S_{0,5}\}$. If $\Sigma = \Sigma' = S_{0,5}$ then the result again holds (using Theorem 15.1). However, if $\Sigma = \Sigma' = S_{1,2}$, then the conclusion of Theorem 15.1 fails without further hypotheses; see [Shackleton 2007]. Note however, that the centre of $\mathrm{Map}(S_{1,2})$ is $\mathbb{Z}_2$, and the quotient is isomorphic to the index-5 subgroup of $\mathrm{Map}(S_{0,5})$ which fixes a boundary curve. Therefore $\mathbb{M}(S_{1,2})$ is quasi-isometric to $\mathbb{M}(S_{0,5})$, and this fact allows us again to describe all quasi-isometric embeddings between the marking complexes of surfaces of complexity 2 up to bounded distance. In particular, they are again all quasi-isometries.

The complexity-1 case corresponds to $S_{1,1}$ or $S_{0,4}$. In these cases the marking complexes are quasitrees, and there are uncountably many classes of quasi-isometries between them up to bounded distance. Finally, the mapping class groups of $S_{0,3}$, $S_{0,2}$, $S_{0,1}$ and $S_{0,0}$ are all finite.

Note that this gives a complete quasi-isometry classification of the groups $\mathrm{Map}(\Sigma)$ — they are all different apart from the classes $\{S_{2,0}, S_{0,6}\}$, $\{S_{1,2}, S_{0,5}\}$, $\{S_{1,1}, S_{1,0}, S_{0,4}\}$ and $\{S_{0,3}, S_{0,2}, S_{0,1}, S_{0,0}\}$.

Acknowledgements

Much of the work for this paper was carried out while visiting the Universities of Toulouse, Grenoble, ETH Zürich, Orsay and the Tokyo Institute of Technology. I am grateful to each of these institutions for their hospitality.

References

[Aougab 2013] T. Aougab, "Uniform hyperbolicity of the graphs of curves", *Geom. Topol.* **17**:5 (2013), 2855–2875. MR Zbl

[Bandelt and Hedlíková 1983] H.-J. Bandelt and J. Hedlíková, "Median algebras", *Discrete Math.* **45**:1 (1983), 1–30. MR Zbl

[Behrstock 2006] J. Behrstock, "Asymptotic geometry of the mapping class group and Teichmüller space", *Geom. Topol.* **10** (2006), 1523–1578. MR Zbl

[Behrstock and Minsky 2008] J. Behrstock and Y. Minsky, "Dimension and rank for mapping class groups", *Ann. of Math.* (2) **167**:3 (2008), 1055–1077. MR Zbl

[Behrstock and Minsky 2011] J. Behrstock and Y. Minsky, "Centroids and the rapid decay property in mapping class groups", *J. Lond. Math. Soc.* (2) **84**:3 (2011), 765–784. MR Zbl

[Behrstock, Druţu and Sapir 2011] J. Behrstock, C. Druţu, and M. Sapir, "Median structures on asymptotic cones and homomorphisms into mapping class groups", *Proc. Lond. Math. Soc.* (3) **102**:3 (2011), 503–554. MR Zbl

[Behrstock, Hagen and Sisto 2015] J. Behrstock, M. F. Hagen, and A. Sisto, "Hierarchically hyperbolic spaces, II: Combination theorems and the distance formula", preprint, 2015. arXiv 1509.00632

[Behrstock, Hagen and Sisto 2017] J. Behrstock, M. F. Hagen, and A. Sisto, "Hierarchically hyperbolic spaces, I: Curve complexes for cubical groups", *Geom. Topol.* **21**:3 (2017), 1731–1804. MR Zbl

[Behrstock, Kleiner, Minsky and Mosher 2012] J. Behrstock, B. Kleiner, Y. Minsky, and L. Mosher, "Geometry and rigidity of mapping class groups", *Geom. Topol.* **16**:2 (2012), 781–888. MR Zbl

[Bowditch 2013] B. H. Bowditch, "Coarse median spaces and groups", *Pacific J. Math.* **261**:1 (2013), 53–93. MR Zbl

[Bowditch 2014a] B. H. Bowditch, "Embedding median algebras in products of trees", *Geom. Dedicata* **170** (2014), 157–176. MR Zbl

[Bowditch 2014b] B. H. Bowditch, "Uniform hyperbolicity of the curve graphs", *Pacific J. Math.* **269**:2 (2014), 269–280. MR Zbl

[Bowditch 2015] B. H. Bowditch, "Large-scale rank and rigidity of the Weil–Petersson metric", preprint, University of Warwick, 2015, Available at http://homepages.warwick.ac.uk/~masgak/preprints.html.

[Bowditch 2016a] B. H. Bowditch, "Large-scale rank and rigidity of the Teichmüller metric", *J. Topol.* **9**:4 (2016), 985–1020. MR Zbl

[Bowditch 2016b] B. H. Bowditch, "Quasi-isometric maps between direct products of hyperbolic spaces", *Internat. J. Algebra Comput.* **26**:4 (2016), 619–633. MR Zbl

[Bowditch 2016c] B. H. Bowditch, "Some properties of median metric spaces", *Groups Geom. Dyn.* **10**:1 (2016), 279–317. MR Zbl

[Chatterji, Druţu and Haglund 2010] I. Chatterji, C. Druţu, and F. Haglund, "Kazhdan and Haagerup properties from the median viewpoint", *Adv. Math.* **225**:2 (2010), 882–921. MR Zbl

[Chepoi 2000] V. Chepoi, "Graphs of some CAT(0) complexes", *Adv. in Appl. Math.* **24**:2 (2000), 125–179. MR Zbl

[Clay, Rafi and Schleimer 2014] M. Clay, K. Rafi, and S. Schleimer, "Uniform hyperbolicity of the curve graph via surgery sequences", *Algebr. Geom. Topol.* **14**:6 (2014), 3325–3344. MR Zbl

[van den Dries and Wilkie 1984] L. van den Dries and A. J. Wilkie, "Gromov's theorem on groups of polynomial growth and elementary logic", *J. Algebra* **89**:2 (1984), 349–374. MR Zbl

[Eilenberg and Steenrod 1952] S. Eilenberg and N. Steenrod, *Foundations of algebraic topology*, Princeton Univ. Press, 1952. MR Zbl

[Eskin, Masur and Rafi 2017] A. Eskin, H. Masur, and K. Rafi, "Large-scale rank of Teichmüller space", *Duke Math. J.* **166**:8 (2017), 1517–1572. MR Zbl

[Farb, Lubotzky and Minsky 2001] B. Farb, A. Lubotzky, and Y. Minsky, "Rank-1 phenomena for mapping class groups", *Duke Math. J.* **106**:3 (2001), 581–597. MR Zbl

[Gromov 1993] M. Gromov, "Asymptotic invariants of infinite groups", pp. 1–295 in *Geometric group theory, II* (Sussex, 1991), edited by G. A. Niblo and M. A. Roller, London Math. Soc. Lecture Note Ser. **182**, Cambridge Univ. Press, 1993. MR Zbl

[Hamenstädt 2005] U. Hamenstädt, "Geometry of the mapping class groups, III: Quasi-isometric rigidity", preprint, 2005. arXiv math/0512429

[Hensel, Przytycki and Webb 2015] S. Hensel, P. Przytycki, and R. C. H. Webb, "1-slim triangles and uniform hyperbolicity for arc graphs and curve graphs", *J. Eur. Math. Soc.* **17**:4 (2015), 755–762. MR Zbl

[Isbell 1980] J. R. Isbell, "Median algebra", *Trans. Amer. Math. Soc.* **260**:2 (1980), 319–362. MR Zbl

[Kapovich, Kleiner and Leeb 1998] M. Kapovich, B. Kleiner, and B. Leeb, "Quasi-isometries and the de Rham decomposition", *Topology* **37**:6 (1998), 1193–1211. MR Zbl

[Kleiner and Leeb 1997] B. Kleiner and B. Leeb, "Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings", *Inst. Hautes Études Sci. Publ. Math.* 86 (1997), 115–197. MR Zbl

[Mangahas 2010] J. Mangahas, "Uniform uniform exponential growth of subgroups of the mapping class group", *Geom. Funct. Anal.* **19**:5 (2010), 1468–1480. MR Zbl

[Masur and Minsky 1999] H. Masur and Y. Minsky, "Geometry of the complex of curves, I: Hyperbolicity", *Invent. Math.* **138**:1 (1999), 103–149. MR Zbl

[Masur and Minsky 2000] H. Masur and Y. Minsky, "Geometry of the complex of curves, II: Hierarchical structure", *Geom. Funct. Anal.* **10**:4 (2000), 902–974. MR Zbl

[Rafi and Schleimer 2011] K. Rafi and S. Schleimer, "Curve complexes are rigid", *Duke Math. J.* **158**:2 (2011), 225–246. MR Zbl

[Roller 1998] M. Roller, *Poc sets, median algebras and group actions: an extended study of Dunwoody's construction and Sageev's theorem*, Habilitationsschrift, Universität Regensburg, 1998. arXiv 1607.07747

[Shackleton 2007] K. J. Shackleton, "Combinatorial rigidity in curve complexes and mapping class groups", *Pacific J. Math.* **230**:1 (2007), 217–232. MR Zbl

[Sholander 1954] M. Sholander, "Medians and betweenness", *Proc. Amer. Math. Soc.* **5** (1954), 801–807. MR Zbl

[Verheul 1993] E. R. Verheul, *Multimedians in metric and normed spaces*, CWI Tract **91**, Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam, 1993. MR Zbl

[Webb 2015] R. C. H. Webb, "Uniform bounds for bounded geodesic image theorems", *J. Reine Angew. Math.* **709** (2015), 219–228. MR Zbl

Received April 8, 2016. Revised January 4, 2017.

BRIAN H. BOWDITCH
MATHEMATICS INSTITUTE
UNIVERSITY OF WARWICK
COVENTRY, CV4 7AL
UNITED KINGDOM

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.75

BACH-FLAT ISOTROPIC GRADIENT RICCI SOLITONS

ESTEBAN CALVIÑO-LOUZAO, EDUARDO GARCÍA-RÍO,
IXCHEL GUTIÉRREZ-RODRÍGUEZ AND RAMÓN VÁZQUEZ-LORENZO

We construct examples of Bach-flat gradient Ricci solitons in neutral signature which are neither half conformally flat nor conformally Einstein.

1. Introduction

Let (M, g) be a pseudo-Riemannian manifold. Let $f \in \mathcal{C}^\infty(M)$. We say that (M, g, f) is a *gradient Ricci soliton* if the equation

$$\mathrm{Hes}_f + \rho = \lambda\, g, \tag{1}$$

is satisfied for some $\lambda \in \mathbb{R}$, where ρ is the *Ricci tensor*, and $\mathrm{Hes}_f = \nabla df$ is the *Hessian tensor* acting on f. A gradient Ricci soliton is said to be *trivial* if the potential function f is constant, since (1) reduces to the Einstein equation $\rho = \lambda g$. Gradient Ricci solitons have been extensively investigated and their classification under geometric conditions is a problem of current interest. We refer to [Cao 2010] for more information.

The gradient Ricci soliton equation codifies geometric information of (M, g) in terms of the Ricci curvature and the second fundamental form of the level sets of the potential function f. The fact that the Ricci tensor completely determines the curvature tensor in the locally conformally flat case has yielded some results in this situation [Cao and Chen 2012; Munteanu and Sesum 2013; Petersen and Wylie 2010]. Any locally conformally flat gradient Ricci soliton is locally a warped product in the Riemannian setting [Fernández-López and García-Río 2014]. The higher signature case, however, allows other possibilities when the level sets of the potential function are degenerate hypersurfaces [Brozos-Vázquez et al. 2013]. Four-dimensional half conformally flat (i.e., self-dual or anti-self-dual) gradient Ricci solitons have been investigated in the Riemannian and neutral signature cases [Brozos-Vázquez and García-Río 2016; Chen and Wang 2015]. While they are locally conformally flat in the Riemannian situation, neutral signature allows other examples given by Riemannian extensions of affine gradient Ricci solitons.

MSC2010: primary 53C25; secondary 53C20, 53C44.
Keywords: gradient Ricci soliton, Bach tensor, Riemannian extension, affine surface.

Let W be the Weyl conformal curvature tensor of (M, g). The Bach tensor, $\mathfrak{B}_{ij} = \nabla^k\nabla^\ell W_{kij\ell} + \frac{1}{2}\rho^{k\ell} W_{kij\ell}$, is conformally invariant in dimension 4. Bach-flat metrics contain half conformally flat and conformally Einstein metrics as special cases [Besse 1987]. Hence, a natural problem is to classify Bach-flat gradient Ricci solitons. The Riemannian case was investigated in [Cao et al. 2014; Cao and Chen 2013] both in the shrinking and steady cases. In all situations the Bach-flat condition reduces to the locally conformally flat one under some natural assumptions.

Our main purpose in this paper is to construct new examples of Bach-flat gradient Ricci solitons in neutral signature. The corresponding potential functions have degenerate level set hypersurfaces and their underlying structure is never locally conformally flat, in sharp contrast with the Riemannian situation. These metrics are realized on the cotangent bundle $T^*\Sigma$ of an affine surface (Σ, D), and they may be viewed as perturbations of the classical Riemannian extensions introduced by Patterson and Walker [1952].

Here is a brief guide to some of the most important results of this paper. In Theorem 3.1 we show that, for any affine surface (Σ, D) admitting a parallel nilpotent $(1, 1)$-tensor field T, the modified Riemannian extension $(T^*\Sigma, g_{D,T,\Phi})$ is Bach-flat. Moreover we show that Bach-flatness is independent of the deformation tensor field Φ, thus providing an infinite family of Bach-flat metrics for any initial data (Σ, D, T). Affine surfaces admitting a parallel nilpotent $(1, 1)$-tensor field T are characterized in Proposition 3.3 by the recurrence of the symmetric part of the Ricci tensor, being $\ker T$ a parallel one-dimensional distribution whose integral curves are geodesics.

The previous construction is used in Theorem 4.3 to show that, for any smooth function $h \in \mathcal{C}^\infty(\Sigma)$, there exist appropriate deformation tensor fields Φ such that $(T^*\Sigma, g_{D,T,\Phi}, f = h \circ \pi)$ is a steady gradient Ricci soliton if and only if $dh(\ker T) = 0$. This provides infinitely many examples of Bach-flat gradient Ricci solitons in neutral signature.

Theorems 5.1 and 6.1 show that $(T^*\Sigma, g_{D,T,\Phi})$ is generically strictly Bach-flat, i.e., neither half conformally flat nor conformally Einstein. Moreover, Theorem 5.1 is used in Proposition 5.2 to construct new examples of anti-self-dual metrics. Turning to gradient Ricci solitons, we show in Theorem 5.4 the existence of anti-self-dual steady gradient Ricci solitons which are not locally conformally flat.

The paper is organized as follows. Some basic results on the Bach tensor and gradient Ricci solitons are introduced in Section 2, as well as a sketch of the construction of modified Riemannian extensions $g_{D,\Phi,T}$. We use these metrics in Section 3 to show that, for any parallel tensor field T on (Σ, D), $g_{D,\Phi,T}$ is Bach-flat if and only if T is either a multiple of the identity or nilpotent (see Theorem 3.1). In Section 4 we show that for each initial data (Σ, D, T) there are an infinite number of Bach-flat steady gradient Ricci solitons (see Theorem 4.3). Nontriviality of the examples is obtained after an examination of the half conformally flat condition (see Section 5) and the conformally Einstein property (see Section 6) of the modified Riemannian extensions introduced in Section 2. Finally, we specialize this construction in Section 7 to provide some illustrative examples.

2. Preliminaries

Let (M^n, g) be a pseudo-Riemannian manifold with Ricci curvature ρ and scalar curvature τ. Let W denote the Weyl conformal curvature tensor and define

$$W[\rho](X,Y) = \sum_{ij} \varepsilon_i \varepsilon_j W(E_i, X, Y, E_j)\rho(E_i, E_j),$$

where $\{E_i\}$ is a local orthonormal frame and $\varepsilon_i = g(E_i, E_i)$. Then the *Bach tensor* is defined (see [Bach 1921]) by

$$\mathfrak{B} = \operatorname{div}_1 \operatorname{div}_4 W + \frac{n-3}{n-2} W[\rho], \tag{2}$$

where div is the divergence operator.

Let $\mathfrak{S} = \rho - \tau/(2(n-1))\, g$ denote the Schouten tensor of (M, g). Let $\mathfrak{C}$ be the *Cotton tensor*, $\mathfrak{C}_{ijk} = (\nabla_i \mathfrak{S})_{jk} - (\nabla_j \mathfrak{S})_{ik}$; it provides a measure of the lack of symmetry on the covariant derivative of the Schouten tensor. Since $\operatorname{div}_4 W = -(n-3)/(n-2)\, \mathfrak{C}$, the Bach and the Cotton tensors of any four-dimensional manifold are related by $\mathfrak{B} = \frac{1}{2}(-\operatorname{div}_1 \mathfrak{C} + W[\rho])$.

The Bach tensor, which is trace-free and conformally invariant in dimension $n = 4$, has been broadly investigated in the literature, both from the geometrical and physical viewpoints (see, for example, [Chen and He 2013; Derdzinski 1983; Dunajski and Tod 2014]). It is the gradient of the L^2 functional of the Weyl curvature on compact manifolds. The field equations of conformal gravity are equivalent to setting the Bach tensor equal to zero and it is also central in the study of the Bach flow, a geometric flow which is quadratic on the curvature and whose fixed points are the vacuum solutions of conformal Weyl gravity [Bakas et al. 2010].

Besides the half conformally flat metrics and the conformally Einstein ones, there are few known examples of strictly Bach-flat manifolds, meaning the ones which are neither half conformally flat nor conformally Einstein (see, for example, [Abbena et al. 2013; Hill and Nurowski 2009; Leistner and Nurowski 2010]). Motivated by this lack of examples, we first construct new explicit four-dimensional Bach-flat manifolds of neutral signature.

Riemannian extensions. In order to introduce the family of metrics under consideration, we recall that a pseudo-Riemannian manifold (M, g) is a *Walker manifold* if there exists a parallel null distribution $\mathcal{D}$ on M. Walker metrics, also called Brinkmann waves in the literature, have been widely investigated in the Lorentzian setting (pp-waves being a special class among them). They appear in many geometrical situations showing a specific behavior without Riemannian counterpart (see [Brozos-Vázquez et al. 2009]).

Let $(M, g, \mathcal{D})$ be a four-dimensional Walker manifold of neutral signature and $\mathcal{D}$ of maximal rank. Then there are local coordinates $(x^1, x^2, x_{1'}, x_{2'})$ so that the metric g is given (see [Walker 1950]) by

$$g = 2\,dx^i \circ dx_{i'} + g_{ij}\,dx^i \circ dx^j, \tag{3}$$

where "$\circ$" denotes the symmetric product $\omega_1 \circ \omega_2 := \frac{1}{2}(\omega_1 \otimes \omega_2 + \omega_2 \otimes \omega_1)$ and (g_{ij}) is a 2×2 symmetric matrix whose entries are functions of all the variables. Moreover, the parallel degenerate distribution is given by $\mathcal{D} = \operatorname{span}\{\partial_{x_{1'}}, \partial_{x_{2'}}\}$.

A special family of four-dimensional Walker metrics is provided by the Riemannian extensions of affine connections to the cotangent bundle of an affine surface. Next we briefly sketch their construction. Let $T^*\Sigma$ be the cotangent bundle of a surface Σ and let $\pi : T^*\Sigma \to \Sigma$ be the projection. Let $\tilde{p} = (p, \omega)$ denote a point of $T^*\Sigma$, where $p \in \Sigma$ and $\omega \in T^*_p\Sigma$. Local coordinates (x^i) in an open set $\mathcal{U}$ of Σ induce local coordinates $(x^i, x_{i'})$ in $\pi^{-1}(\mathcal{U})$, where one sets $\omega = \sum x_{i'}\,dx^i$. The evaluation functions on $T^*\Sigma$ play a central role in the construction. They are defined as follows. For each vector field X on Σ, *the evaluation of* X is the real valued function $\iota X : T^*\Sigma \to \mathbb{R}$ given by $\iota X(p, \omega) = \omega(X_p)$. Vector fields on $T^*\Sigma$ are characterized by their action on evaluations ιX and one defines the complete lift to $T^*\Sigma$ of a vector field X on Σ by $X^C(\iota Z) = \iota[X, Z]$, for all vector fields Z on Σ. Moreover, a $(0, s)$-tensor field on $T^*\Sigma$ is characterized by its action on complete lifts of vector fields on Σ.

Next, let D be a torsion-free affine connection on Σ. The *Riemannian extension* g_D is the neutral signature metric g_D on $T^*\Sigma$ characterized by the identity $g_D(X^C, Y^C) = -\iota(D_X Y + D_Y X)$ (see [Patterson and Walker 1952]). They are expressed in the induced local coordinates $(x^i, x_{i'})$ as follows:

$$g_D = 2\,dx^i \circ dx_{i'} - 2x_{k'}{}^D\Gamma_{ij}{}^k dx^i \circ dx^j, \tag{4}$$

where ${}^D\Gamma_{ij}{}^k$ denote the Christoffel symbols of D. The geometry of $(T^*\Sigma, g_D)$ is strongly related to that of (Σ, D). Recall that the curvature of any affine surface is completely determined by its Ricci tensor ρ^D. Moreover, the symmetric and skew-symmetric parts given by $\rho^D_{\mathrm{sym}}(X, Y) = \frac{1}{2}\{\rho^D(X, Y) + \rho^D(Y, X)\}$ and $\rho^D_{\mathrm{sk}}(X, Y) = \frac{1}{2}\{\rho^D(X, Y) - \rho^D(Y, X)\}$ play a distinguished role.

Let Φ be a symmetric $(0, 2)$-tensor field on Σ. Then the *deformed Riemannian extension*, $g_{D,\Phi} = g_D + \pi^*\Phi$, is a first perturbation of the Riemannian extension. A second one is obtained as follows. Let $T = T_i^k dx^i \otimes \partial_{x^k}$ be a $(1, 1)$-tensor field on Σ. Its evaluation ιT defines a one-form on $T^*\Sigma$ characterized by $\iota T(X^C) = \iota(TX)$. The *modified Riemannian extension* $g_{D,\Phi,T}$ is the neutral signature metric on $T^*\Sigma$ defined (see [Calviño-Louzao et al. 2009]) by

$$g_{D,\Phi,T} = \iota T \circ \iota T + g_D + \pi^*\Phi, \tag{5}$$

where Φ is a symmetric $(0,2)$-tensor field on Σ. In local coordinates one has

$$g_{D,\Phi,T} = 2\,dx^i \circ dx_{i'} + \left\{\tfrac{1}{2}x_{r'}x_{s'}(T_i^r T_j^s + T_j^r T_i^s) - 2x_{k'}\,{}^D\Gamma_{ij}{}^k + \Phi_{ij}\right\} dx^i \circ dx^j.$$

The case when T is a multiple of the identity ($T = c\,\mathrm{Id}$, $c \neq 0$) is of special interest. It was shown in [Calviño-Louzao et al. 2009] that for any affine surface (Σ, D), the modified Riemannian extension $g_{D,\Phi,c\,\mathrm{Id}}$ is an Einstein metric on $T^*\Sigma$ if and only if the deformation tensor Φ is the symmetric part of the Ricci tensor of (Σ, D). Moreover, a slight generalization of the modified Riemannian extension allows a complete description of self-dual Walker metrics as follows.

Theorem 2.1 [Calviño-Louzao et al. 2009; Díaz-Ramos et al. 2006]. A four-dimensional Walker metric is self-dual if and only if it is locally isometric to the cotangent bundle $T^*\Sigma$ of an affine surface (Σ, D), with metric tensor

$$g = \iota X(\iota\,\mathrm{id} \circ \iota\,\mathrm{id}) + \iota\,\mathrm{id} \circ \iota T + g_D + \pi^*\Phi,$$

where X, T, D and Φ are a vector field, a $(1,1)$-tensor field, a torsion-free affine connection and a symmetric $(0,2)$-tensor field on Σ, respectively.

As a matter of notation, we will write $\partial_k = \partial/\partial x^k$ and $\partial_{k'} = \partial/\partial x_{k'}$, unless we want to emphasize some special coordinates. We will let

$$\phi_k = (\partial/\partial x^k)\phi \quad \text{and} \quad \phi_{k'} = (\partial/\partial x_{k'})\phi$$

denote the corresponding first derivatives of a smooth function ϕ.

Gradient Ricci solitons and affine gradient Ricci solitons. Let (M, g, f) be a gradient Ricci soliton. The level set hypersurfaces of the potential function play a distinguished role in analyzing the geometry of gradient Ricci solitons. Hence we say that the soliton is *nonisotropic* if ∇f is nowhere lightlike (i.e., $\|\nabla f\|^2 \neq 0$), and that the soliton is *isotropic* if $\|\nabla f\|^2 = 0$, but $\nabla f \neq 0$.

Nonisotropic gradient Ricci solitons lead to local warped product decompositions in the locally conformally flat and half conformally flat cases, and their geometry resembles the Riemannian situation [Brozos-Vázquez and García-Río 2016; Brozos-Vázquez et al. 2013]. The isotropic case is, however, in sharp contrast with the positive definite setting since ∇f gives rise to a Walker structure. Self-dual gradient Ricci solitons which are not locally conformally flat are isotropic and steady ($\lambda = 0$ in (1)). Moreover, they are described in terms of Riemannian extensions as follows.

Theorem 2.2 [Brozos-Vázquez and García-Río 2016]. Let (M, g, f) be a four-dimensional self-dual gradient Ricci soliton of neutral signature which is not locally conformally flat. Then (M, g) is locally isometric to the cotangent bundle $T^*\Sigma$ of an affine surface (Σ, D) equipped with a modified Riemannian extension $g_{D,\Phi,0}$.

Moreover any such gradient Ricci soliton is steady and the potential function is given by $f = h \circ \pi$ for some $h \in \mathcal{C}^\infty(\Sigma)$ satisfying the affine gradient Ricci soliton equation

$$\mathrm{Hes}_h^D + 2\rho_{\mathrm{sym}}^D = 0, \tag{6}$$

for any symmetric (0, 2)-tensor field Φ on Σ.

The previous result relates the affine geometry of (Σ, D) and the pseudo-Riemannian geometry of $(T^*\Sigma, g_{D,\Phi,0})$, allowing the construction of an infinite family of steady gradient Ricci solitons on $T^*\Sigma$ for any initial data (Σ, D, h) satisfying (6). It is important to remark here that the existence of affine gradient Ricci solitons imposes some restrictions on (Σ, D), as shown in [Brozos-Vázquez et al. 2018] in the locally homogeneous case.

3. Bach-flat modified Riemannian extensions

The use of modified Riemannian extensions with $T = c\,\mathrm{Id}$ allowed the construction of many examples of self-dual Einstein metrics [Calviño-Louzao et al. 2009]. One of the crucial facts in understanding the metrics $g_{D,\Phi,c\,\mathrm{Id}}$ is that the (1, 1)-tensor field $T = c\,\mathrm{Id}$ is parallel with respect to the connection D. Hence, a natural generalization arises by considering arbitrary tensor fields T which are parallel with respect to the affine connection D.

Let (Σ, D, T) be a torsion-free affine surface equipped with a parallel (1, 1)-tensor field T. Parallelizability of T guarantees the existence of local coordinates (x^1, x^2) on Σ so that

$$T\partial_1 = T_1^1\,\partial_1 + T_1^2\,\partial_2,$$

and

$$T\partial_2 = T_2^1\,\partial_1 + T_2^2\,\partial_2,$$

for some real constants T_i^j. Let $(T^*\Sigma, g_{D,\Phi,T})$ be the modified Riemannian extension given by (5). Further note that D and Φ are taken with full generality. Thus, the corresponding Christoffel symbols ${}^D\Gamma_{ij}^k$ and the coefficient functions Φ_{ij} are arbitrary smooth functions of the coordinates (x^1, x^2).

Our first main result concerns the construction of Bach-flat metrics:

Theorem 3.1. Let (Σ, D, T) be a torsion-free affine surface equipped with a parallel (1, 1)-tensor field T. Let Φ be an arbitrary symmetric (0, 2)-tensor field on Σ. Then the Bach tensor of $(T^*\Sigma, g_{D,\Phi,T})$ vanishes if and only if T is either a multiple of the identity or nilpotent.

Proof. In order to compute the Bach tensor of $(T^*\Sigma, g_{D,\Phi,T})$, first of all observe that being T-parallel imposes some restrictions on the components T_i^j as well as on the Christoffel symbols of the connection D:

$$
DT = 0 \Rightarrow \begin{cases}
T_2^1\, {}^D\Gamma_{11}^2 - T_1^2\, {}^D\Gamma_{12}^1 = 0, \\
T_2^1\, {}^D\Gamma_{12}^2 - T_1^2\, {}^D\Gamma_{22}^1 = 0, \\
T_1^2\, {}^D\Gamma_{11}^1 + (T_2^2 - T_1^1)\, {}^D\Gamma_{11}^2 - T_1^2\, {}^D\Gamma_{12}^2 = 0, \\
T_2^1\, {}^D\Gamma_{11}^1 + (T_2^2 - T_1^1)\, {}^D\Gamma_{12}^1 - T_2^1\, {}^D\Gamma_{12}^2 = 0, \\
T_1^2\, {}^D\Gamma_{12}^1 + (T_2^2 - T_1^1)\, {}^D\Gamma_{12}^2 - T_1^2\, {}^D\Gamma_{22}^2 = 0, \\
T_2^1\, {}^D\Gamma_{12}^1 + (T_2^2 - T_1^1)\, {}^D\Gamma_{22}^1 - T_2^1\, {}^D\Gamma_{22}^2 = 0.
\end{cases} \tag{7}
$$

Then, expressing the Bach tensor $\mathfrak{B}_{ij} = \mathfrak{B}(\partial_i, \partial_j)$ in induced coordinates $(x^i, x_{i'})$, a long but straightforward calculation shows that

$$
(\mathfrak{B}_{ij}) = \left(\begin{array}{cc|c} \mathfrak{B}_{11} & \mathfrak{B}_{12} & \multirow{2}{*}{$\widetilde{\mathfrak{B}}$} \\ \mathfrak{B}_{12} & \mathfrak{B}_{22} & \\ \hline \multicolumn{2}{c|}{\widetilde{\mathfrak{B}}} & 0 \end{array} \right), \tag{8}
$$

where

$$
\widetilde{\mathfrak{B}} = \tfrac{1}{6}\big((T_1^1 - T_2^2)^2 + 4T_2^1 T_1^2\big) \cdot (T_1^1 + T_2^2) \cdot \begin{pmatrix} T_1^1 - T_2^2 & 2T_1^2 \\ 2T_2^1 & T_2^2 - T_1^1 \end{pmatrix}
$$

and where the coefficients $\mathfrak{B}_{11}$, $\mathfrak{B}_{12}$ and $\mathfrak{B}_{22}$ can be written in terms of $\mathfrak{d} = \det(T)$ and $\mathfrak{t} = \operatorname{tr}(T)$ as follows:

$$
\begin{aligned}
\mathfrak{B}_{11} = &-\tfrac{1}{6}\big\{10\mathfrak{d}^3 - 2\big(\mathfrak{t}^2 + 13T_2^2\mathfrak{t} - 15(T_2^2)^2\big)\mathfrak{d}^2 + (5\mathfrak{t} - T_2^2)(\mathfrak{t} - T_2^2)\mathfrak{t}^2\mathfrak{d} - (\mathfrak{t} - T_2^2)^2\mathfrak{t}^4\big\} x_{1'}^2 \\
&- \tfrac{1}{6}\big\{(T_1^2)^2(30\mathfrak{d}^2 + \mathfrak{t}^2\mathfrak{d} - \mathfrak{t}^4)\big\} x_{2'}^2 \\
&- \tfrac{1}{3}\big\{(13\mathfrak{t} - 30T_2^2)\mathfrak{d}^2 + (3\mathfrak{t} - T_2^2)\mathfrak{t}^2\mathfrak{d} - (\mathfrak{t} - T_2^2)\mathfrak{t}^4\big\} T_1^2\, x_{1'} x_{2'} \\
&- \tfrac{1}{3}\big\{\big({}^D\Gamma_{11}^1 + 2\, {}^D\Gamma_{12}^2\big)(\mathfrak{t} - 2T_2^2) + 2T_1^2\, {}^D\Gamma_{22}^2\big\}(\mathfrak{t}^2 - 4\mathfrak{d})\mathfrak{t}\, x_{1'} \\
&- \tfrac{1}{3}\big\{{}^D\Gamma_{11}^2(\mathfrak{t} - 2T_2^2) + 2T_1^2\, {}^D\Gamma_{12}^2\big\}(\mathfrak{t}^2 - 4\mathfrak{d})\mathfrak{t}\, x_{2'} \\
&- \tfrac{1}{6}\big\{10\mathfrak{d}^2 + \big(3\mathfrak{t}^2 - 22T_2^2\mathfrak{t} + 14(T_2^2)^2\big)\mathfrak{d} - \big(\mathfrak{t}^2 - 4T_2^2\mathfrak{t} + 2(T_2^2)^2\big)\mathfrak{t}^2\big\}\Phi_{11} \\
&- \tfrac{1}{3}\big\{(11\mathfrak{t} - 14T_2^2)\mathfrak{d} - 2(\mathfrak{t} - T_2^2)\mathfrak{t}^2\big\} T_1^2 \Phi_{12} \\
&+ \tfrac{1}{3}\big\{\mathfrak{t}^2 - 7\mathfrak{d}\big\}(T_1^2)^2 \Phi_{22} \\
&- \tfrac{2}{3}\big(\partial_2\, {}^D\Gamma_{11}^2 - \partial_1\, {}^D\Gamma_{12}^2\big)(4\mathfrak{d} - \mathfrak{t}^2),
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{B}_{12} = -\tfrac{1}{6}\big\{&(13\mathfrak{t}-30T_2^2)\mathfrak{d}^2+(3\mathfrak{t}-T_2^2)\mathfrak{t}^2\mathfrak{d}-(\mathfrak{t}-T_2^2)\mathfrak{t}^4\big\}T_2^1x_{1'}^2\\
&+\tfrac{1}{6}\big\{(17\mathfrak{t}-30T_2^2)\mathfrak{d}^2-(2\mathfrak{t}+T_2^2)\mathfrak{t}^2\mathfrak{d}+T_2^2\mathfrak{t}^4\big\}T_1^2x_{2'}^2\\
&+\tfrac{1}{6}\big\{20\mathfrak{d}^3+4\big(4\mathfrak{t}^2-15T_2^2\mathfrak{t}+15(T_2^2)^2\big)\mathfrak{d}^2\\
&\qquad-\big(3\mathfrak{t}^2+2T_2^2\mathfrak{t}-2(T_2^2)^2\big)\mathfrak{t}^2\mathfrak{d}+2(\mathfrak{t}-T_2^2)T_2^2\mathfrak{t}^4\big\}x_{1'}x_{2'}\\
&-\tfrac{1}{3}\big\{{}^D\Gamma_{12}^1(\mathfrak{t}-2T_2^2)+2T_1^2\,{}^D\Gamma_{22}^1\big\}(\mathfrak{t}^2-4\mathfrak{d})\mathfrak{t}x_{1'}\\
&-\tfrac{1}{3}\big\{{}^D\Gamma_{12}^2(\mathfrak{t}-2T_2^2)+2T_1^2\,{}^D\Gamma_{22}^2\big\}(\mathfrak{t}^2-4\mathfrak{d})\mathfrak{t}x_{2'}\\
&-\tfrac{1}{6}\big\{(11\mathfrak{t}-14T_2^2)\mathfrak{d}-2(\mathfrak{t}-T_2^2)\mathfrak{t}^2\big\}T_2^1\Phi_{11}\\
&+\tfrac{1}{6}\big\{4\mathfrak{d}^2+\big(6\mathfrak{t}^2-28T_2^2\mathfrak{t}+28(T_2^2)^2\big)\mathfrak{d}-(\mathfrak{t}-2T_2^2)^2\mathfrak{t}^2\big\}\Phi_{12}\\
&+\tfrac{1}{6}\big\{(3\mathfrak{t}-14T_2^2)\mathfrak{d}+2T_2^2\mathfrak{t}^2\big\}T_1^2\Phi_{22}\\
&-\tfrac{1}{3}\big\{\big(\partial_2\,{}^D\Gamma_{11}^1-\partial_1\,{}^D\Gamma_{12}^1-\partial_2\,{}^D\Gamma_{12}^2+\partial_1\,{}^D\Gamma_{22}^2\big)(\mathfrak{t}^2-4\mathfrak{d})\big\},
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{B}_{22} = &-\tfrac{1}{6}\big\{30\mathfrak{d}^2-\mathfrak{t}^4+\mathfrak{t}^2\mathfrak{d}\big\}(T_2^1)^2x_{1'}^2\\
&-\tfrac{1}{6}\big\{10\mathfrak{d}^3+2\big(\mathfrak{t}^2-17T_2^2\mathfrak{t}+15(T_2^2)^2\big)\mathfrak{d}^2+(4\mathfrak{t}+T_2^2)T_2^2\mathfrak{t}^2\mathfrak{d}-(T_2^2)^2\mathfrak{t}^4\big\}x_{2'}^2\\
&+\tfrac{1}{3}\big\{(17\mathfrak{t}-30T_2^2)\mathfrak{d}^2-(2\mathfrak{t}+T_2^2)\mathfrak{t}^2\mathfrak{d}+T_2^2\mathfrak{t}^4\big\}T_2^1x_{1'}x_{2'}\\
&-\tfrac{1}{3}\big\{{}^D\Gamma_{22}^1(\mathfrak{t}-2T_2^2)+2T_2^1\,{}^D\Gamma_{22}^2\big\}(\mathfrak{t}^2-4\mathfrak{d})\mathfrak{t}x_{1'}\\
&+\tfrac{1}{3}\big\{{}^D\Gamma_{22}^2(\mathfrak{t}-2T_2^2)-2T_1^2\,{}^D\Gamma_{22}^1\big\}(\mathfrak{t}^2-4\mathfrak{d})\mathfrak{t}x_{2'}\\
&-\tfrac{1}{3}(7\mathfrak{d}-\mathfrak{t}^2)(T_2^1)^2\Phi_{11}\\
&+\tfrac{1}{3}\big\{(3\mathfrak{t}-14T_2^2)T_2^1\mathfrak{d}+2T_2^1T_2^2\mathfrak{t}^2\big\}\Phi_{12}\\
&-\tfrac{1}{6}\big\{10\mathfrak{d}^2-\big(5\mathfrak{t}^2+6T_2^2\mathfrak{t}-14(T_2^2)^2\big)\mathfrak{d}+\mathfrak{t}^4-2(T_2^2)^2\mathfrak{t}^2\big\}\Phi_{22}\\
&-\tfrac{2}{3}(\partial_2\,{}^D\Gamma_{12}^1-\partial_1\,{}^D\Gamma_{22}^1)(\mathfrak{t}^2-4\mathfrak{d}).
\end{aligned}
$$

Suppose first that the Bach tensor of $(T^*\Sigma, g_{\tilde{D},\Phi,T})$ vanishes. We start analyzing the case $T_2^1=0$. In this case, the expression of $\widetilde{\mathfrak{B}}$ in (8) reduces to

$$
\widetilde{\mathfrak{B}} = \tfrac{1}{6}(T_1^1-T_2^2)^2\cdot(T_1^1+T_2^2)\cdot\begin{pmatrix} T_1^1-T_2^2 & 2T_1^2\\ 0 & T_2^2-T_1^1\end{pmatrix}. \tag{9}
$$

If $T_2^2=T_1^1$, we differentiate the component $\mathfrak{B}_{11}$ in (8) twice with respect to $x_{2'}$ to obtain $T_1^2T_1^1=0$. Thus, either $T_1^2=0$ and T is a multiple of the identity, or $T_1^1=0$ and, in such a case, T is determined by $T\partial_1=T_1^2\partial_2$ and therefore it is nilpotent. If $T_2^2\neq T_1^1$, then (9) implies that $T_2^2=-T_1^1$. In this case, we differentiate the component $\mathfrak{B}_{22}$ in (8) twice with respect to $x_{2'}$ and obtain $T_1^1 = 0$. Thus, as before, T is nilpotent.

Next we analyze the case $T_2^1 \neq 0$. We use (7) to express

$$
{}^D\Gamma_{11}^1 = \frac{T_1^1 - T_2^2}{T_2^1}\, {}^D\Gamma_{12}^1 + \frac{T_1^2}{T_2^1}\, {}^D\Gamma_{22}^1, \qquad {}^D\Gamma_{11}^2 = \frac{T_1^2}{T_2^1}\, {}^D\Gamma_{12}^1,
$$
$$
{}^D\Gamma_{12}^2 = \frac{T_1^2}{T_2^1}\, {}^D\Gamma_{22}^1, \qquad {}^D\Gamma_{22}^2 = {}^D\Gamma_{12}^1 - \frac{T_1^1 - T_2^2}{T_2^1}\, {}^D\Gamma_{22}^1.
$$

Considering the component $\widetilde{\mathfrak{B}}_{11}$ in (8),

$$
\widetilde{\mathfrak{B}}_{11} = \tfrac{1}{6}(T_1^1 - T_2^2)\cdot(T_1^1 + T_2^2)\cdot\big((T_1^1 - T_2^2)^2 + 4T_2^1 T_1^2\big),
$$

we analyze separately the vanishing of each one of the three factors in $\widetilde{\mathfrak{B}}_{11}$.

Assume that $T_2^2 = T_1^1$. In this case, component $\widetilde{\mathfrak{B}}_{12}$ in (8) reduces to $\widetilde{\mathfrak{B}}_{12} = \frac{8}{3}T_2^1(T_1^2)^2 T_1^1$; since we are assuming that $T_2^1 \neq 0$, then either $T_1^2 = 0$ or $T_1^2 \neq 0$ and $T_1^1 = 0$. If $T_1^2 = 0$, the only nonzero component of the Bach tensor is given by

$$
\mathfrak{B}_{22} = -(T_2^1)^2 (T_1^1)^2 (3(T_1^1)^2 x_{1'}^2 + \Phi_{11}),
$$

from where it follows that $T_1^1 = 0$ and hence T is determined by $T\partial_2 = T_2^1\partial_1$ and is nilpotent. If $T_1^2 \neq 0$ and $T_1^1 = 0$, then we differentiate the component $\mathfrak{B}_{12}$ in (8) with respect to $x_{1'}$ and $x_{2'}$ to get $T_2^1 T_1^2 = 0$, which is not possible since both T_2^1 and T_1^2 are non-null.

Suppose now that $T_2^2 = -T_1^1$. In this case, we differentiate the component $\mathfrak{B}_{22}$ in (8) twice with respect to $x_{1'}$ and as a consequence we obtain $T_2^1(T_2^1 T_1^2 + (T_1^1)^2) = 0$; since we are assuming $T_2^1 \neq 0$, it follows that $T_1^2 = -(T_1^1)^2/T_2^1$. Thus, the (1, 1)-tensor field T is given by $T\partial_1 = T_1^1\partial_1 - (T_1^1)^2/T_2^1\,\partial_2$ and $T\partial_2 = T_2^1\partial_1 - T_1^1\partial_2$, and therefore it is nilpotent as well.

Finally, suppose that $(T_1^1 - T_2^2)^2 + 4T_2^1 T_1^2 = 0$; since $T_2^1 \neq 0$, this is equivalent to $T_1^2 = -(T_1^1 - T_2^2)^2/(4T_2^1)$. Now, we differentiate the component $\mathfrak{B}_{22}$ in (8) twice with respect to $x_{1'}$ to obtain $T_2^1(T_1^1 + T_2^2) = 0$. Thus, we have $T_2^2 = -T_1^1$ and T is given by $T\partial_1 = T_1^1\partial_1 - (T_1^1)^2/T_2^1\,\partial_2$ and $T\partial_2 = T_2^1\partial_1 - T_1^1\partial_2$, which again implies that T is nilpotent.

To conclude the proof we show the "only if" part. If T is a multiple of the identity, then $(T^*\Sigma, g_{D,\Phi,T})$ is self-dual by Theorem 2.1 and therefore it has vanishing Bach tensor. Thus, we suppose T is parallel and nilpotent and, in this case, we can choose a system of coordinates (x^1, x^2) such that T is determined by $T\partial_1 = \partial_2$ and $T\partial_2 = 0$. Hence, examining (8), clearly $\widetilde{\mathfrak{B}} = 0$ and, since $\mathfrak{d} = \mathfrak{t} = 0$, one easily checks that $\mathfrak{B}_{11} = \mathfrak{B}_{12} = \mathfrak{B}_{22} = 0$, showing that the Bach tensor of $(T^*\Sigma, g_{D,\Phi,T})$ vanishes. $\square$

Remark 3.2. We emphasize that even though the Bach tensor of the metrics $g_{D,\Phi,T}$ depends on the choice of Φ (as shown in the proof of Theorem 3.1), the existence of Bach-flat metrics in Theorem 3.1 is independent of the symmetric $(0,2)$-tensor field Φ, thus providing an infinite family of examples for each initial data (Σ, D, T). Moreover, note that the metrics $g_{D,\Phi,T}$ are generically nonisometric for different deformation tensor fields Φ.

The Bach-flat modified Riemannian extensions in Theorem 3.1 obtained from a $(1,1)$-tensor field of the form $T = c\,\mathrm{Id}$ are not of interest for our purposes since they all are half conformally flat (see Theorem 2.1). Hence, in what follows we focus on the case when T is a parallel nilpotent $(1,1)$-tensor field and refer to $g_{D,\Phi,T}$ as a *nilpotent Riemannian extension*.

Affine connections supporting parallel nilpotent tensors. It is shown in the proof of Theorem 3.1 that the existence of a parallel nilpotent tensor field T on a torsion-free affine surface (Σ, D) imposes some restrictions on D.

Proposition 3.3. Let (Σ, D, T) be a torsion-free affine surface equipped with a nilpotent $(1,1)$-tensor field T. If T is parallel, then:

(i) $\ker T$ is a parallel one-dimensional distribution whose integral curves are geodesics of (Σ, D).

(ii) The symmetric part of the Ricci tensor, ρ^D_{sym}, is zero or of rank one and recurrent, i.e.,
$$D\rho^D_{\mathrm{sym}} = \eta \otimes \rho^D_{\mathrm{sym}},$$
for some one-form η.

Proof. Let (Σ, D) be a torsion-free affine surface admitting a parallel nilpotent $(1,1)$-tensor field T. Then there exist suitable coordinates (x^1, x^2) where $T\partial_1 = \partial_2$, $T\partial_2 = 0$ and it follows from (7) that the Christoffel symbols of D satisfy

$$(10)\qquad {}^D\Gamma_{12}^1 = 0, \qquad {}^D\Gamma_{12}^2 = {}^D\Gamma_{11}^1, \qquad {}^D\Gamma_{22}^1 = 0, \qquad {}^D\Gamma_{22}^2 = 0.$$

In such a case the one-dimensional distribution $\ker T (= \mathrm{span}\{\partial_2\})$ is parallel and ∂_2 is a geodesic vector field, thus showing (i). Moreover, the Ricci tensor of any affine connection given by (10) satisfies

$$\rho^D = \begin{pmatrix} \partial_2\, {}^D\Gamma_{11}^2 - \partial_1\, {}^D\Gamma_{11}^1 & \partial_2\, {}^D\Gamma_{11}^1 \\ -\partial_2\, {}^D\Gamma_{11}^1 & 0 \end{pmatrix},$$

from where it follows that the symmetric and the skew-symmetric parts of the Ricci tensor are given by

$$\rho^D_{\mathrm{sym}} = \left(\partial_2\, {}^D\Gamma_{11}^2 - \partial_1\, {}^D\Gamma_{11}^1\right) dx^1 \circ dx^1, \qquad \rho^D_{\mathrm{sk}} = \partial_2\, {}^D\Gamma_{11}^1\, dx^1 \wedge dx^2.$$

Hence ρ^D_{sym} is either zero or of rank one. Moreover, a straightforward calculation of the covariant derivative of the symmetric part of the Ricci tensor gives

$$(D_{\partial_1}\rho^D_{\mathrm{sym}})(\partial_1,\partial_1) = \partial_{12}\,{}^D\Gamma_{11}^2 - \partial_{11}\,{}^D\Gamma_{11}^1 - 2\,{}^D\Gamma_{11}^1(\partial_2\,{}^D\Gamma_{11}^2 - \partial_1\,{}^D\Gamma_{11}^1),$$
$$(D_{\partial_2}\rho^D_{\mathrm{sym}})(\partial_1,\partial_1) = \partial_{22}\,{}^D\Gamma_{11}^2 - \partial_{12}\,{}^D\Gamma_{11}^1,$$

with the other components being zero. This shows that ρ^D_{sym} is recurrent, i.e., $D\rho^D_{\mathrm{sym}} = \eta\otimes\rho^D_{\mathrm{sym}}$, with recurrence one-form

$$\eta = \left\{\partial_1 \ln \rho^D_{\mathrm{sym}}(\partial_1,\partial_1) - 2\,{}^D\Gamma_{11}^1\right\} dx^1 + \partial_2 \ln \rho^D_{\mathrm{sym}}(\partial_1,\partial_1)\, dx^2, \tag{11}$$

which proves (ii). □

Remark 3.4. It follows from the expression of ρ^D_{sk} in the proof of Proposition 3.3 that any connection given by (10) has symmetric Ricci tensor if and only if $\partial_2\,{}^D\Gamma_{11}^1 = 0$, in which case ρ^D is recurrent. Now, it follows from the work of Wong [1964] that any such connection can be described in suitable coordinates $(\bar u^1,\bar u^2)$ by

$$D_{\partial_{\bar u^1}}\partial_{\bar u^1} = {}^{\bar u}\Gamma_{11}^2(\bar u^1,\bar u^2)\partial_{\bar u^2},$$

where ${}^{\bar u}\Gamma_{11}^2(\bar u^1,\bar u^2)$ is an arbitrary function satisfying $\partial_{\bar u^2}\,{}^{\bar u}\Gamma_{11}^2(\bar u^1,\bar u^2)\neq 0$. Further, the only nonzero component of the Ricci tensor is $\rho^D(\partial_{\bar u^1},\partial_{\bar u^1}) = \partial_{\bar u^2}\,{}^{\bar u}\Gamma_{11}^2$, and the recurrence one-form ω is given by

$$\omega = \partial_{\bar u^1}(\ln \partial_{\bar u^2}\,{}^{\bar u}\Gamma_{11}^2)d\bar u^1 + \partial_{\bar u^2}(\ln \partial_{\bar u^2}\,{}^{\bar u}\Gamma_{11}^2)d\bar u^2. \tag{12}$$

Further assume that T is a parallel nilpotent $(1,1)$-tensor field on (Σ, D). Then a straightforward calculation shows that its expression in the coordinates $(\bar u^1,\bar u^2)$ is given by $T\partial_{\bar u^1} = T_1^2\,\partial_{\bar u^2}$ and $T\partial_{\bar u^2} = 0$, for some $T_1^2\in\mathbb{R}$, $T_1^2\neq 0$. Hence, considering the modified coordinates $(u^1,u^2) = (\bar u^1, (T_1^2)^{-1}\bar u^2)$ one has that $T\partial_{u^1} = \partial_{u^2}$ and $T\partial_{u^2} = 0$, and the connection is determined by the only nonzero Christoffel symbol ${}^u\Gamma_{11}^2$. Moreover, it follows from the expression of the recurrence one-form ω that $\omega(\ker T) = 0$ if and only if $\partial_{22}\,{}^u\Gamma_{11}^2 = 0$.

4. Bach-flat gradient Ricci solitons

Let Φ be a symmetric $(0,2)$-tensor field on (Σ, D, T). One uses the nilpotent structure T to construct an associated symmetric $(0,2)$-tensor field $\widehat\Phi$ given by $\widehat\Phi(X,Y) = \Phi(TX,TY)$, for all vector fields X, Y on Σ. Further, let (x^1,x^2) be local coordinates where $T\partial_1 = \partial_2$, $T\partial_2 = 0$ and let $\Phi = \Phi_{ij}dx^i\otimes dx^j$. Then $\widehat\Phi$ expresses as $\widehat\Phi = \widehat\Phi_{ij}dx^i\otimes dx^j = \Phi_{22}dx^1\otimes dx^1$.

Einstein nilpotent Riemannian extensions.

Theorem 4.1. Let (Σ, D, T) be an affine surface equipped with a parallel nilpotent $(1,1)$-tensor field T and let Φ be a symmetric $(0,2)$-tensor field on Σ. Then $(T^*\Sigma, g_{D,\Phi,T})$ is Einstein (indeed, Ricci-flat) if and only if $\widehat{\Phi} = -2\rho^D_{\mathrm{sym}}$.

Proof. Let (x^1, x^2) be local coordinates on Σ so that $T\partial_1 = \partial_2$, $T\partial_2 = 0$, and consider the induced coordinates $(x^1, x^2, x_{1'}, x_{2'})$ on $T^*\Sigma$. A straightforward calculation shows that the Ricci tensor of any nilpotent Riemannian extension $g_{D,\Phi,T}$ is determined by

$$\rho(\partial_1, \partial_1) = \Phi(\partial_2, \partial_2) + 2\rho^D_{\mathrm{sym}}(\partial_1, \partial_1),$$

the other components being zero. Hence the Ricci operator is nilpotent and $g_{D,\Phi,T}$ has zero scalar curvature. Moreover, the Ricci tensor vanishes if and only if $\Phi(\partial_2, \partial_2) + 2\rho^D_{\mathrm{sym}}(\partial_1, \partial_1) = 0$. The result now follows. □

Remark 4.2. The Weyl tensor of a pseudo-Riemannian manifold is harmonic if and only if the Cotton tensor vanishes. Let (Σ, D, T) be an affine surface equipped with a parallel nilpotent $(1,1)$-tensor field T and let Φ be a symmetric $(0,2)$-tensor field on Σ. Let (x^1, x^2) be local coordinates on Σ so that $T\partial_1 = \partial_2$, $T\partial_2 = 0$, and consider the induced coordinates $(x^1, x^2, x_{1'}, x_{2'})$ on $T^*\Sigma$. A straightforward calculation shows that the Cotton tensor of $(T^*\Sigma, g_{D,\Phi,T})$ is given by

$$\mathfrak{C}(\partial_1, \partial_2, \partial_1) = -\{\partial_2\, \Phi(\partial_2, \partial_2) + 2\partial_2\, \rho^D_{\mathrm{sym}}(\partial_1, \partial_1)\},$$

the other components being zero. Hence $(T^*\Sigma, g_{D,\Phi,T})$ *has harmonic Weyl tensor if and only if* $\widehat{D\Phi} = -2\,\widehat{\eta} \otimes \rho^D_{\mathrm{sym}}$, *where* $\widehat{\eta}(X) = \eta(TX)$, η *being the recurrence one-form given in* (11), *and* $\widehat{D\Phi}(X, Y; Z) = D\Phi(TX, TY; TZ)$.

Gradient Ricci solitons on nilpotent Riemannian extensions. From Theorem 2.2, recall that the affine gradient Ricci soliton equation $\mathrm{Hes}^D_h + 2\rho^D_{\mathrm{sym}} = 0$ determines the potential function of any self-dual gradient Ricci soliton which is not locally conformally flat, independently of the deformation tensor Φ. The next theorem shows that, in contrast with the previous situation, for any $h \in \mathcal{C}^\infty(\Sigma)$ with $dh(\ker T) = 0$, one may use the symmetric $(0,2)$-tensor field $\mathrm{Hes}^D_h + 2\rho^D_{\mathrm{sym}}$ to determine a deformation tensor field Φ so that the resulting nilpotent Riemannian extension is a Bach-flat steady gradient Ricci soliton with potential function $f = h \circ \pi$.

Theorem 4.3. Let (Σ, D, T) be an affine surface equipped with a parallel nilpotent $(1,1)$-tensor field T and let Φ be a symmetric $(0,2)$-tensor field on Σ. Let $h \in \mathcal{C}^\infty(\Sigma)$ be a smooth function. Then $(T^*\Sigma, g_{D,\Phi,T}, f = h \circ \pi)$ is a Bach-flat gradient Ricci soliton if and only if $dh(\ker T) = 0$ and

$$\widehat{\Phi} = -\,\mathrm{Hes}^D_h - 2\rho^D_{\mathrm{sym}}. \tag{13}$$

Moreover the soliton is steady and isotropic.

Proof. Let (x^1, x^2) be local coordinates on Σ so that $T\partial_1 = \partial_2$, $T\partial_2 = 0$, and consider the induced coordinates $(x^1, x^2, x_{1'}, x_{2'})$ on $T^*\Sigma$. Setting $f = h \circ \pi$, one has that $\mathrm{Hes}_f(\partial_1, \partial_{1'}) + \rho(\partial_1, \partial_{1'}) = \lambda g(\partial_1, \partial_{1'})$ leads to $\lambda = 0$, which shows that the soliton is steady. A straightforward calculation shows that the remaining nonzero terms in the gradient Ricci soliton equation are given by

$$\begin{aligned}
&\mathrm{Hes}_f(\partial_2, \partial_2) + \rho(\partial_2, \partial_2) = \partial_{22}h,\\
&\mathrm{Hes}_f(\partial_1, \partial_2) + \rho(\partial_1, \partial_2) = \partial_{12}h - {}^D\Gamma_{11}^1 \partial_2 h,\\
&\mathrm{Hes}_f(\partial_1, \partial_1) + \rho(\partial_1, \partial_1) = x_{2'}\,\partial_2 h - {}^D\Gamma_{11}^2\,\partial_2 h + \partial_{11}h - {}^D\Gamma_{11}^1 \partial_1 h\\
&\qquad\qquad\qquad\qquad\qquad\qquad + \Phi_{22} + 2\partial_2\,{}^D\Gamma_{11}^2 - 2\partial_1\,{}^D\Gamma_{11}^1.
\end{aligned}$$

It immediately follows from the equation $(\mathrm{Hes}_f + \rho)(\partial_1, \partial_1) = 0$ that $\partial_2 h = 0$, which shows that $dh(\ker T) = 0$. The only remaining equation now becomes

$$\begin{aligned}
\mathrm{Hes}_f(\partial_1, \partial_1) + \rho(\partial_1, \partial_1) &= \partial_{11}h - {}^D\Gamma_{11}^1 \partial_1 h + \Phi_{22} + 2\partial_2\,{}^D\Gamma_{11}^2 - 2\partial_1\,{}^D\Gamma_{11}^1\\
&= \Phi(\partial_2, \partial_2) + \mathrm{Hes}_h^D(\partial_1, \partial_1) + 2\rho_{\mathrm{sym}}^D(\partial_1, \partial_1),
\end{aligned}$$

from which (13) follows. Moreover, it also follows from the form of the potential function that $\nabla f = h'(x^1)\partial_{1'}$, and thus $\|\nabla f\|^2 = 0$ (equivalently the level hypersurfaces of the potential function are degenerate submanifolds of $T^*\Sigma$), which shows that the soliton is isotropic. □

Remark 4.4. The potential functions of the gradient Ricci solitons in Theorem 4.3 are of the form $f = h \circ \pi$ for some $h \in \mathcal{C}^\infty(\Sigma)$. Next we show that this is indeed the case if the Ricci tensor of (Σ, D) is nonsymmetric.

Let $(T^*\Sigma, g_{D,\Phi,T}, f)$ be a gradient Ricci soliton with potential function $f \in \mathcal{C}^\infty(T^*\Sigma)$. Take local coordinates $(x^1, x^2, x_{1'}, x_{2'})$ on $T^*\Sigma$ as in the proof of Theorem 4.3. Since $\mathrm{Hes}_f(\partial_{i'}, \partial_{j'}) = \partial_{i'j'} f(x^1, x^2, x_{1'}, x_{2'})$, it follows from the expression of the Ricci tensor in Theorem 4.1 and the metric tensor (5), that the potential function is determined by $f = \iota X + h \circ \pi$, for some $h \in \mathcal{C}^\infty(\Sigma)$ and some vector field X on Σ, where ιX is the evaluation map acting on X.

Further set $X = A(x^1, x^2)\partial_1 + B(x^1, x^2)\partial_2$ in the local coordinates (x^1, x^2) above, for some $A, B \in \mathcal{C}^\infty(\Sigma)$. Then $\mathrm{Hes}_f(\partial_2, \partial_{1'}) = \partial_2 A(x^1, x^2)$, from where it follows that $X = A(x^1)\partial_1 + B(x^1, x^2)\partial_2$. Considering the component $\mathrm{Hes}_f(\partial_2, \partial_{2'}) = -A''(x^1) + \partial_2 B(x^1, x^2)$, one has that $X = A(x^1)\partial_1 + (P(x^1) + x^2 A'(x^1))\partial_2$ for some smooth function $P(x^1)$. Next the component

$$\begin{aligned}
\mathrm{Hes}_f(\partial_1, \partial_{2'}) &= A(x^1)\,{}^D\Gamma_{11}^2 - x_{2'} A(x^1)\\
&\quad + {}^D\Gamma_{11}^1\big(P(x^1) + x^2 A'(x^1)\big) + P'(x^1) + x^2 A''(x^1)
\end{aligned}$$

shows that $A = 0$ and it reduces to $\mathrm{Hes}_f(\partial_1, \partial_{2'}) = P'(x^1) + P(x^1)\,{}^D\Gamma_{11}^1$. A solution $P(x^1)$ of the equation $P'(x^1) + P(x^1)\,{}^D\Gamma_{11}^1 = 0$ either vanishes identically (and hence $X = 0$) or it is nowhere zero, in which case $\partial_2{}^D\Gamma_{11}^1 = 0$ (see the proof of Theorem 6.1). In the latter case Proposition 3.3 shows that the Ricci tensor of (Σ, D) is symmetric and thus recurrent of rank one. Therefore Theorem 4.3 describes all possible gradient Ricci solitons on $(T^*\Sigma, g_{D,\Phi,T})$ whenever ρ_{sk}^D is nonzero.

Remark 4.5. The tensor field $\mathbb{D}_{ijk} = \mathfrak{C}_{ijk} + W_{ijk\ell}\nabla_\ell f$ introduced in [Cao and Chen 2013] plays an essential role in analyzing the geometry of Bach-flat gradient Ricci solitons. Local conformal flatness in [Cao et al. 2014; Cao and Chen 2013] follows from $\mathbb{D} = 0$, which is obtained under some natural assumptions.

Gradient Ricci solitons in Theorem 4.3 satisfy $\nabla f = h'(x^1)\partial_{1'}$. Then, a straightforward calculation shows that $\mathbb{D}$ is completely determined by

$$\mathbb{D}_{121} = -2h'(x^1)\partial_2{}^D\Gamma_{11}^1(x^1, x^2),$$

the other components being zero. Hence it follows from the proof of Proposition 3.3 that the tensor field $\mathbb{D}$ vanishes if and only if the Ricci tensor ρ^D is symmetric. However Theorem 5.1 shows that $(T^*\Sigma, g_{D,\Phi,T})$ is never locally conformally flat.

5. Half conformally flat nilpotent Riemannian extensions

The existence of a null distribution $\mathcal{D}$ on a four-dimensional manifold (M, g) of neutral signature defines a natural orientation on M: the one which, for any basis u, v of $\mathcal{D}$, makes the bivector $u \wedge v$ self-dual (see [Derdzinski 2008]). We consider on $T^*\Sigma$ the orientation which agrees with $\mathcal{D} = \ker \pi_*$, and thus self-duality and anti-self-duality are not interchangeable. The following result shows that they are essentially different for nilpotent Riemannian extensions.

Theorem 5.1. Let (Σ, D, T) be an affine surface equipped with a parallel nilpotent $(1, 1)$-tensor field T. Then

(i) $(T^*\Sigma, g_{D,\Phi,T})$ is never self-dual for any deformation tensor field Φ.

(ii) If $(T^*\Sigma, g_{D,\Phi,T})$ is anti-self-dual, then D is either a flat connection or (Σ, D) is recurrent with symmetric Ricci tensor of rank one.

In the later case there exist local coordinates (u^1, u^2) where the only nonzero Christoffel symbol is ${}^u\Gamma_{11}^2$ and the tensor field T is given by $T\partial_{u^1} = \partial_{u^2}$, $T\partial_{u^2} = 0$. Moreover, $(T^*\Sigma, g_{D,\Phi,T})$ is anti-self-dual if and only if the symmetric $(0, 2)$-tensor field Φ satisfies the equations

$$\begin{aligned} &\widehat{D\Phi} = -2\widehat{\omega} \otimes \rho^D, \\ &0 = \tfrac{1}{2}\widehat{\Phi} \otimes \widehat{\Phi}(\partial_1, \partial_1, \partial_1, \partial_1) + 2(\widehat{\Phi} \otimes \rho^D)(\partial_1, \partial_1, \partial_1, \partial_1) \\ &\qquad + D^2\Phi(\partial_1, \partial_1; T\partial_1, T\partial_1) + D^2\Phi(T\partial_1, T\partial_1; \partial_1, \partial_1) \\ &\qquad - 2\, D^2\Phi(\partial_1, T\partial_1; T\partial_1, \partial_1), \end{aligned} \tag{14}$$

where

$$\widehat{D\Phi}(X, Y, Z) = D\Phi(TX, TY; TZ),$$

ω is the recurrence one-form given by $D\rho^D = \omega \otimes \rho^D$, and $\widehat{\omega}(X) = \omega(TX)$.

Proof. A direct computation using the expression of the anti-self-dual curvature operator of any four-dimensional Walker metric obtained in [Díaz-Ramos et al. 2006] shows that, for any nilpotent Riemannian extension $g_{D,\Phi,T}$, W^- takes the form

$$W^- = \frac{1}{2}\begin{pmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \tag{15}$$

thus showing that the anti-self-dual Weyl curvature operator W^- is nilpotent and hence $(T^*\Sigma, g_{D,\Phi,T})$ is never self-dual, which proves (i).

Next we show (ii). Let (M, g) be a four-dimensional Walker metric (3) and set the metric components $g_{11} = a$, $g_{12} = c$ and $g_{22} = b$, where g_{ij} are functions of the Walker coordinates $(x^1, x^2, x_{1'}, x_{2'})$. Then the self-dual Weyl curvature operator takes the form (see [Díaz-Ramos et al. 2006])

$$W^+ = \begin{pmatrix} W_{11}^+ & W_{12}^+ & W_{11}^+ + \frac{\tau}{12} \\ -W_{12}^+ & \frac{\tau}{6} & -W_{12}^+ \\ -W_{11}^+ - \frac{\tau}{12} & -W_{12}^+ & -W_{11}^+ - \frac{\tau}{6} \end{pmatrix}, \tag{16}$$

where

$$\begin{aligned} W_{11}^+ = \tfrac{1}{12}\big(&6ca_1b_2 - 6a_1b_{1'} - 6ba_1c_2 + 12a_1c_{2'} - 6ca_2b_1 + 6a_2b_{2'} \\ &+ 6ba_2c_1 + 6a_{1'}b_1 - 6a_{2'}b_2 - 12a_{2'}c_1 + 6ab_1c_2 - 6ab_2c_1 \\ &+ 12b_2c_{1'} - 12b_{1'}c_2 - a_{11} - 12c^2a_{11} - 12bca_{12} + 24ca_{12'} \\ &- 3b^2a_{22} + 12ba_{22'} - 12a_{2'2'} - 3a^2b_{11} + 12ab_{11'} - b_{22} \\ &- 12b_{1'1'} + 12acc_{11} - 2c_{12} + 6abc_{12} - 24cc_{11'} - 12ac_{12'} \\ &- 12bc_{21'} + 24c_{1'2'}\big), \end{aligned} \tag{17}$$

and

$$\begin{aligned} W_{12}^+ = \tfrac{1}{4}\big(&-2ca_{11} - ba_{12} + 2a_{12'} + ab_{12} - 2b_{21'} + ac_{11} - 2cc_{12} \\ &-2c_{11'} - bc_{22} + 2c_{22'}\big). \end{aligned} \tag{18}$$

Since any anti-self-dual metric is Bach-flat, we proceed as in the proof of Theorem 3.1 considering local coordinates (x^1, x^2) on the surface Σ such that T is determined by $T\partial_1 = \partial_2$ and $T\partial_2 = 0$. Since T is parallel, the Christoffel symbols must satisfy (10), i.e.,

$$ {}^D\Gamma_{12}^1=0, \qquad {}^D\Gamma_{12}^2={}^D\Gamma_{11}^1, \qquad {}^D\Gamma_{22}^1=0, \qquad {}^D\Gamma_{22}^2=0. $$

Next, we analyze the self-dual Weyl curvature operator, which is completely determined by the scalar curvature and its components W_{11}^+ and W_{12}^+ already described in equations (17) and (18). The scalar curvature is zero by Theorem 4.1, and $W_{12}^+=-2\partial_2{}^D\Gamma_{11}^1$, from where it follows that the Ricci tensor ρ^D is symmetric of rank one and recurrent (see Remark 3.4). Take local coordinates (u^1,u^2) as in Remark 3.4 so that the only nonzero Christoffel symbol is ${}^u\Gamma_{11}^2$ and $T\partial_{u^1}=\partial_{u^2}$, $T\partial_{u^2}=0$. Finally, we compute the component W_{11}^+ given by (17) in the coordinates $(u^1,u^2,u_{1'},u_{2'})$ of $T^*\Sigma$, obtaining

$$ \begin{aligned} W_{11}^+ = (\partial_2\Phi_{22}+2\partial_{22}{}^u\Gamma_{11}^2)\,u_{2'} - \tfrac{1}{2}(\Phi_{22})^2 - 2\Phi_{22}\partial_2{}^u\Gamma_{11}^2 - \partial_2\Phi_{22}{}^u\Gamma_{11}^2 \\ +2\partial_{12}\Phi_{12}-\partial_{22}\Phi_{11}-\partial_{11}\Phi_{22}. \end{aligned} $$

Thus $(T^*\Sigma, g_{D,\Phi,T})$ is anti-self-dual if and only if

$$ \begin{aligned} \partial_2\Phi_{22}+2\partial_{22}{}^u\Gamma_{11}^2 &= 0, \\ \tfrac{1}{2}(\Phi_{22})^2+2\Phi_{22}\partial_2{}^u\Gamma_{11}^2+\partial_2\Phi_{22}{}^u\Gamma_{11}^2 &= 2\partial_{12}\Phi_{12}-\partial_{22}\Phi_{11}-\partial_{11}\Phi_{22}, \end{aligned} $$

from where (14) follows. □

Anti-self-dual gradient Ricci solitons. Self-dual gradient Ricci solitons which are not locally conformally flat are described in Theorem 2.2. In contrast, no explicit examples of strictly anti-self-dual gradient Ricci solitons were previously reported. In this section we use nilpotent Riemannian extensions to construct anti-self-dual isotropic gradient Ricci solitons. In this case, Theorem 5.1 shows that (Σ, D) must have symmetric Ricci tensor.

Proposition 5.2. Let (Σ, D, T, Φ) be an affine surface with symmetric Ricci tensor equipped with a parallel nilpotent $(1,1)$-tensor field T and a parallel symmetric $(0,2)$-tensor field Φ. Then $(T^*\Sigma, g_{D,\Phi,T})$ is anti-self-dual if and only if $\widehat{\omega}=0$ and $\widehat{\Phi}=0$, where ω is the recurrence one-form given by (12).

Proof. If the Ricci tensor ρ^D is symmetric of rank one and Φ is parallel, then the equations in Theorem 5.1 reduce to $\widehat{\omega}=0$ and $\widehat{\Phi}=0$, which proves the result. If (Σ, D) is a flat surface then a straightforward calculation shows that anti-self-duality is equivalent to $\widehat{\Phi}=0$, being Φ a parallel tensor. □

Since the deformation tensor Φ of any gradient Ricci soliton in Theorem 4.3 must satisfy $\widehat{\Phi}=-\operatorname{Hes}_h^D-2\rho_{\mathrm{sym}}^D$, the condition $\widehat{\Phi}=0$ in the previous proposition restricts the consideration of Ricci solitons on $(T^*\Sigma, g_{D,\Phi,T})$ to those originated by affine gradient Ricci solitons on (Σ, D).

Proposition 5.3. Let (Σ, D, T) be an affine surface equipped with a parallel nilpotent $(1, 1)$-tensor field T and let $h \in \mathcal{C}^\infty(\Sigma)$. Then:

(i) (Σ, D, T, h) is an affine gradient Ricci soliton with $dh(\ker T) = 0$ if and only if $(T^*\Sigma, g_{D,\widehat{\Phi},T}, f = h \circ \pi)$ is a Bach-flat steady gradient Ricci soliton for any symmetric $(0, 2)$-tensor field Φ.

(ii) (Σ, D, T, h) is a nonflat affine gradient Ricci soliton with $dh(\ker T) = 0$ if and only if the recurrence one-form η in (11) satisfies $\widehat{\eta} = 0$.

Proof. Since T is nilpotent, $\widehat{\Phi}(TX, TY) = 0$ for any $(0, 2)$-tensor field Φ. Hence (13) shows that $(T^*\Sigma, g_{D,\widehat{\Phi},T}, f = h \circ \pi)$ is a gradient Ricci soliton if and only if (Σ, D, T, h) is an affine gradient Ricci soliton with $dh(\ker T) = 0$, which shows (i).

Next take local coordinates (x^1, x^2) on Σ so that $T\partial_1 = \partial_2$, $T\partial_2 = 0$. Since the Christoffel symbols ${}^D\Gamma_{ij}^k$ are given by (10), using the expression of ρ^D_{sym} in Proposition 3.3, one has $(\mathrm{Hes}^D_h + 2\rho^D_{\mathrm{sym}})(\partial_2, \partial_2) = \partial_{22}h$. Thus $h(x^1, x^2) = x^2 P(x^1) + Q(x^1)$ for some $P, Q \in \mathcal{C}^\infty(\Sigma)$. Hence $dh(\ker T) = 0$ holds if and only if $P = 0$. Since $h(x^1, x^2) = Q(x^1)$, one has that $(\mathrm{Hes}^D_h + 2\rho^D_{\mathrm{sym}})(\partial_1, \partial_2) = 0$, and the only remaining equation is

$$0 = (\mathrm{Hes}^D_h + 2\rho^D_{\mathrm{sym}})(\partial_1, \partial_1) = Q'' + 2\big(\partial_2{}^D\Gamma_{11}^2 - \partial_1{}^D\Gamma_{11}^1\big) = Q'' + 2\rho^D(\partial_1, \partial_1).$$

Therefore, the integrability condition becomes $\partial_2\rho^D(\partial_1, \partial_1) = 0$. Hence, it follows from (11) that (Σ, D, T, h) is an affine gradient Ricci soliton with $dh(\ker T) = 0$ if and only if the symmetric part of the Ricci tensor ρ^D_{sym} is recurrent with recurrence one-form η satisfying $\eta(\ker T) = 0$. Assertion (ii) now follows. $\square$

A direct application of the previous propositions gives the desired examples.

Theorem 5.4. Let (Σ, D, T, Φ) be an affine surface with symmetric Ricci tensor equipped with a parallel nilpotent $(1, 1)$-tensor field T and a parallel symmetric $(0, 2)$-tensor field Φ.

(i) (Σ, D, h) is an affine gradient Ricci soliton with $dh(\ker T) = 0$ if and only if $(T^*\Sigma, g_{D,\widehat{\Phi},T}, f = h \circ \pi)$ is an anti-self-dual steady gradient Ricci soliton which is not locally conformally flat.

(ii) (Σ, D, h) is an affine gradient Ricci soliton with $dh(\ker T) = 0$ if and only if there exist local coordinates (u^1, u^2) on Σ so that the only nonzero Christoffel symbol is given by ${}^u\Gamma_{11}^2 = P(u^1) + u^2 Q(u^1)$ and the potential function $h(u^1)$ is determined by $h''(u^1) = -2Q(u^1)$, for any $P, Q \in \mathcal{C}^\infty(\Sigma)$.

Proof. $(T^*\Sigma, g_{D,\widehat{\Phi},T}, f = h \circ \pi)$ is a gradient Ricci soliton by Proposition 5.3(i). Anti-self-duality now follows from Proposition 5.2 and Proposition 5.3(ii), showing assertion (i).

Assertion (ii) follows from Proposition 5.3(ii) and the expression of the recurrence form ω in (12). Take local coordinates (u^1, u^2) on Σ as in the proof of

Proposition 5.3(ii). Then it follows from (12) that $\widehat{\omega}=0$ if and only if $\partial_{22}{}^u\Gamma_{11}^2=0$. Thus,

$$^u\Gamma_{11}^2(u^1,u^2)=P(u^1)+u^2Q(u^1),$$

for some $P, Q\in\mathcal{C}^\infty(\Sigma)$ and $h''(u^1)=-2Q(u^1)$. □

6. Conformally Einstein nilpotent Riemannian extensions

A pseudo-Riemannian manifold (M^n, g) is said to be (*locally*) *conformally Einstein* if every point $p\in M$ has an open neighborhood $\mathcal{U}$ and a positive smooth function φ defined on $\mathcal{U}$ such that $(\mathcal{U}, \bar g=\varphi^{-2}g)$ is Einstein. Brinkmann [1924] showed that a manifold is conformally Einstein if and only if the equation

$$(n-2)\,\mathrm{Hes}_\varphi+\varphi\,\rho-\tfrac{1}{n}\{(n-2)\Delta\varphi+\varphi\,\tau\}g=0 \tag{19}$$

has a positive solution. Despite its apparent simplicity, the integration of the conformally Einstein equation is surprisingly difficult (see [Kühnel and Rademacher 2008] for more information). It was shown in [Gover and Nagy 2007; Kozameh et al. 1985] that any four-dimensional conformally Einstein manifold satisfies

$$\text{(i)}\quad \mathfrak{C}+W(\cdot,\cdot,\cdot,\nabla\sigma)=0,\qquad\qquad \text{(ii)}\quad \mathfrak{B}=0, \tag{20}$$

where the conformal metric is given by $\bar g=e^{2\sigma}g$.

Conditions (i) and (ii) above are also sufficient to be conformally Einstein if (M, g) is *weakly-generic* (i.e., the Weyl tensor viewed as a map $TM\to\bigotimes^3 TM$ is injective). Since nilpotent Riemannian extensions are not weakly generic (see the expression of W^- in the proof of Theorem 5.1), we will analyze the conformally Einstein equation (19), seeking solutions on nilpotent Riemannian extensions $(T^*\Sigma, g_{D,\Phi,T})$.

Theorem 6.1. Let (Σ, D, T) be a torsion-free affine surface equipped with a parallel nilpotent $(1,1)$-tensor field T. Then any solution of (19) is of the form $\varphi=\iota X+\phi\circ\pi$ for some vector field X on Σ such that $X\in\ker T$ and $\mathrm{tr}(DX)=0$.

Moreover $(T^*\Sigma, g_{D,\Phi,T})$ is conformally Einstein if and only if one of the following holds:

(i) The conformally Einstein equation (19) admits a solution $\varphi=\phi\circ\pi$ for some $\phi\in\mathcal{C}^\infty(\Sigma)$ with $d\phi(\ker T)=0$, and the deformation tensor Φ is determined by

$$\phi\,\widehat{\Phi}+2(\mathrm{Hes}_\phi^D+\phi\,\rho_{\mathrm{sym}}^D)=0.$$

(ii) The conformally Einstein equation (19) admits a solution $\varphi=\iota X+\phi\circ\pi$ for some $\phi\in\mathcal{C}^\infty(\Sigma)$ and some nonzero vector field X on Σ such that $X\in\ker T$ and $\mathrm{tr}(DX)=0$.

In this case, the Ricci tensor ρ^D is symmetric of rank one and recurrent. Moreover, there are local coordinates (u^1, u^2) on Σ so that

$$\varphi(u^1, u^2, u_{1'}, u_{2'}) = \kappa u_{2'} + \phi(u^1, u^2)$$

is a solution of (19) if and only if

$$d\phi(T\partial_1) = \tfrac{\kappa}{2}\Phi(T\partial_1, T\partial_1),$$

$$\begin{aligned}\mathrm{Hes}^D_\phi(\partial_1, \partial_1) + \phi\,\rho^D(\partial_1, \partial_1) &= -\tfrac{1}{2}\big(\phi + 2\kappa\,{}^u\Gamma_{11}^2\big)\Phi(T\partial_1, T\partial_1)\\ &\quad + \tfrac{\kappa}{2}\big\{2(D_{\partial_1}\Phi)(T\partial_1, \partial_1) - (D_{T\partial_1}\Phi)(\partial_1, \partial_1)\big\}.\end{aligned}$$

Proof. Let (x^1, x^2) be local coordinates on Σ so that $T\partial_1 = \partial_2$, $T\partial_2 = 0$, and consider the induced coordinates $(x^1, x^2, x_{1'}, x_{2'})$ on $T^*\Sigma$. Since T is parallel, we obtain directly from (7) that

$${}^D\Gamma_{12}^1 = 0, \quad {}^D\Gamma_{12}^2 = {}^D\Gamma_{11}^1, \quad {}^D\Gamma_{22}^1 = 0, \quad {}^D\Gamma_{22}^2 = 0.$$

In order to analyze the conformally Einstein equation (19), consider the symmetric (0, 2)-tensor field

$$\mathcal{E} = 2\,\mathrm{Hes}_\varphi + \varphi\,\rho - \tfrac{1}{4}\{2\Delta\varphi + \varphi\,\tau\}g$$

and set $\mathcal{E} = 0$. Let $\mathcal{E}_{ij} = \mathcal{E}(\partial_i, \partial_j)$ and let $\varphi \in \mathcal{C}^\infty(T^*\Sigma)$ be a solution of (19). Then one computes

$$\mathcal{E}_{33} = 2\partial_{1'1'}\varphi, \quad \mathcal{E}_{34} = 2\partial_{1'2'}\varphi, \quad \mathcal{E}_{44} = 2\partial_{2'2'}\varphi,$$

to show that any solution of (19) must be of the form

$$\varphi(x^1, x^2, x_{1'}, x_{2'}) = A(x^1, x^2)x_{1'} + B(x^1, x^2)x_{2'} + \psi(x^1, x^2), \tag{21}$$

for some smooth functions A, B and ψ depending only on the coordinates (x^1, x^2). This shows that any solution of the conformally Einstein equation on $(T^*\Sigma, g_{D,\Phi,T})$ is of the form

$$\varphi = \iota X + \psi \circ \pi,$$

where ιX is the evaluation of a vector field $X = A\partial_1 + B\partial_2$ on Σ, $\psi \in \mathcal{C}^\infty(\Sigma)$ and $\pi : T^*\Sigma \to \Sigma$ is the projection.

Now, the conformally Einstein condition given in (19) can be expressed in matrix form as follows:

$$(\mathcal{E}_{ij}) = \begin{pmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} & \partial_1 A - \partial_2 B & 2\big({}^D\Gamma_{11}^2 A + {}^D\Gamma_{11}^1 B + \partial_1 B - A x_{2'}\big) \\ * & \mathcal{E}_{22} & 2\partial_2 A & -\partial_1 A + \partial_2 B \\ * & * & 0 & 0 \\ * & * & * & 0 \end{pmatrix} \tag{22}$$

where positions with $*$ are not written since the matrix is symmetric, and where

$$
\begin{aligned}
\mathcal{E}_{11} = &-\big(\partial_1 A - \partial_2 B - 4\,{}^D\Gamma_{11}^1 A\big)x_{2'}^2 \\
&+\big\{A\Phi_{22} + 2\big(\partial_{11}A - {}^D\Gamma_{11}^2\partial_2 A + {}^D\Gamma_{11}^1\partial_2 B + A\partial_2{}^D\Gamma_{11}^2 - B\partial_2{}^D\Gamma_{11}^1\big)\big\}x_{1'} \\
&-\big\{B\Phi_{22} + 2A\Phi_{12} - 2\big(\partial_{11}B + {}^D\Gamma_{11}^2\partial_1 A - {}^D\Gamma_{11}^1\partial_1 B \\
&\quad + \big(\partial_1{}^D\Gamma_{11}^2 - 2\,{}^D\Gamma_{11}^1\,{}^D\Gamma_{11}^2\big)A + \big(\partial_1{}^D\Gamma_{11}^1 - 2\big({}^D\Gamma_{11}^1\big)^2\big)B + \partial_2\psi\big)\big\}x_{2'} \\
&+ 2\partial_2 A x_{1'}x_{2'} \\
&- (\partial_1 A + \partial_2 B)\Phi_{11} + 2\big({}^D\Gamma_{11}^2 A + {}^D\Gamma_{11}^1 B\big)\Phi_{12} + \big(2\,{}^D\Gamma_{11}^2 B + \psi\big)\Phi_{22} \\
&- A\partial_1\Phi_{11} + B\partial_2\Phi_{11} - 2B\partial_1\Phi_{12} \\
&+ 2\partial_{11}\psi - 2\,{}^D\Gamma_{11}^1\partial_1\psi - 2\,{}^D\Gamma_{11}^2\partial_2\psi - 2(\partial_1{}^D\Gamma_{11}^1 - \partial_2{}^D\Gamma_{11}^2)\psi,
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{E}_{12} = &\,2\big(\partial_{12}A - {}^D\Gamma_{11}^1\partial_2 A + A\partial_2{}^D\Gamma_{11}^1\big)x_{1'} \\
&+ 2\big(\partial_{12}B + {}^D\Gamma_{11}^1\partial_1 A + A\partial_2{}^D\Gamma_{11}^2\big)x_{2'} \\
&- (\partial_1 A + \partial_2 B)\Phi_{12} + 2\,{}^D\Gamma_{11}^1 B\Phi_{22} - A\partial_2\Phi_{11} - B\partial_1\Phi_{22} \\
&+ 2\partial_{12}\psi - 2\,{}^D\Gamma_{11}^1\partial_2\psi,
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{E}_{22} = &\,2\partial_{22}A x_{1'} + 2\big(\partial_{22}B + 2A\partial_2{}^D\Gamma_{11}^1\big)x_{2'} \\
&-\big(\partial_1 A + \partial_2 B + 2\,{}^D\Gamma_{11}^1 A\big)\Phi_{22} - 2A\partial_2\Phi_{12} + A\partial_1\Phi_{22} - B\partial_2\Phi_{22} + 2\partial_{22}\psi.
\end{aligned}
$$

First, we use component

$$\mathcal{E}_{14} = 2\big({}^D\Gamma_{11}^2 A + {}^D\Gamma_{11}^1 B + \partial_1 B - A x_{2'}\big)$$

in (22); note that $\partial_{2'}\mathcal{E}_{14} = -2A$, and so $A(x^1, x^2) = 0$, which shows that $X \in \ker T$. Now component $\mathcal{E}_{13}$ in (22) gives $\partial_2 B = 0$, which implies $B(x^1, x^2) = P(x^1)$ for some smooth function P depending only on the coordinate x^1, i.e., the vector field $X = B\partial_2$ satisfies $\operatorname{tr}(DX) = 0$.

At this point, the conformal function φ has the coordinate expression

$$\varphi(x^1, x^2, x_{1'}, x_{2'}) = P(x^1)x_{2'} + \psi(x^1, x^2)$$

and the possible nonzero components in (22) are $\mathcal{E}_{11}$, $\mathcal{E}_{12}$, $\mathcal{E}_{22}$ and $\mathcal{E}_{14}$. Considering the component $\mathcal{E}_{14} = 2(P'(x^1) + {}^D\Gamma_{11}^1(x^1, x^2)P(x^1))$, we distinguish two cases depending on whether the function P vanishes identically or not. Indeed, if $P(x^1)$ is a solution of the equation $\mathcal{E}_{14} = 0$, then

$$\partial_1\big(P(x^1)e^{\int {}^D\Gamma_{11}^1(x^1,x^2)\,dx^1}\big) = e^{\int {}^D\Gamma_{11}^1(x^1,x^2)\,dx^1}\big\{P'(x^1) + P(x^1)\,{}^D\Gamma_{11}^1(x^1, x^2)\big\} = 0,$$

which shows that $P(x^1)e^{\int {}^D\Gamma_{11}^1(x^1,x^2)\,dx^1} = \mathcal{Q}(x^2)$ for some smooth function $\mathcal{Q}(x^2)$. Now, if the function $\mathcal{Q}(x^2)$ vanishes at some point, then $P(x^1) = 0$ at each point. Otherwise, if $\mathcal{Q}(x^2)$ is not equal to 0 at each point, neither is $P(x^1)$.

First, suppose that $P(x^1) \equiv 0$, and hence $\varphi = \psi \circ \pi$. In this case, component $\mathcal{E}_{22}$ in (22) yields $\partial_{22}\psi = 0$, which implies $\psi(x^1, x^2) = Q(x^1)x^2 + \phi(x^1)$ for some smooth functions Q and ϕ depending only on the coordinate x^1. Now, the only components in (22) which could be non-null are

$$\begin{aligned}\mathcal{E}_{11} &= 2Qx_{2'} + \big(Q\Phi_{22} + 2Q'' - 2\,{}^D\Gamma_{11}^1 Q' - 2\big(\partial_1 {}^D\Gamma_{11}^1 - \partial_2 {}^D\Gamma_{11}^2\big)Q\big)x_2 \\ &\qquad + \phi\Phi_{22} + 2\phi'' - 2\,{}^D\Gamma_{11}^1\phi' - 2\big(\partial_1 {}^D\Gamma_{11}^1 - \partial_2 {}^D\Gamma_{11}^2\big)\phi - 2\,{}^D\Gamma_{11}^2 Q, \\ \mathcal{E}_{12} &= 2\big(Q' - {}^D\Gamma_{11}^1 Q\big).\end{aligned}$$

Now, $\partial_{2'}\mathcal{E}_{11} = 2Q$ implies $Q = 0$, thus showing that $d\varphi(\ker T) = 0$. Then $\mathcal{E}_{12} = 0$ and the component $\mathcal{E}_{11}$ reduces to

$$\mathcal{E}_{11} = \phi\Phi_{22} + 2\phi'' - 2\,{}^D\Gamma_{11}^1\phi' - 2\big(\partial_1 {}^D\Gamma_{11}^1 - \partial_2 {}^D\Gamma_{11}^2\big)\phi.$$

Since $\varphi(x^1, x^2, x_{1'}, x_{2'}) = \phi(x^1)$, ϕ must be non-null and we obtain that $\mathcal{E}_{11} = 0$ is equivalent to

$$\begin{aligned}\Phi_{22} &= -\tfrac{2}{\phi}\big\{\phi'' - {}^D\Gamma_{11}^1\phi' - \big(\partial_1 {}^D\Gamma_{11}^1 - \partial_2 {}^D\Gamma_{11}^2\big)\,\phi\big\}, \\ &= -\tfrac{2}{\phi}\big\{\mathrm{Hes}_\phi^D(\partial_1, \partial_1) + \phi\,\rho_{\mathrm{sym}}^D(\partial_1, \partial_1)\big\},\end{aligned}$$

from where (i) is obtained.

Finally, we analyze the case in which the function $P(x^1)$ does not vanish identically. Since $\mathcal{E}_{14} = 2(P'(x^1) + {}^D\Gamma_{11}^1(x^1, x^2)P(x^1))$, we have $\partial_2 {}^D\Gamma_{11}^1 = 0$. Now it follows from Remark 3.4 that the Ricci tensor ρ^D is symmetric of rank one and recurrent. Specialize the local coordinates (u^1, u^2) on Σ so that the only nonzero Christoffel symbol of D is ${}^u\Gamma_{11}^2(u^1, u^2)$ and $T\partial_{u^1} = \partial_{u^2}$, $T\partial_{u^2} = 0$. Then any solution of the conformally Einstein equation takes the form

$$\varphi(u^1, u^2, u_{1'}, u_{2'}) = \mathcal{A}(u^1)u_{2'} + \phi(u^1, u^2).$$

Now, considering the component $\mathcal{E}_{41}$ of the conformally Einstein equation in the new coordinates (u^1, u^2), one has $\mathcal{E}_{41} = 2\,\mathcal{A}'(u^1)$, which shows that $\varphi(u^1, u^2, u_{1'}, u_{2'}) = \kappa u_{2'} + \phi(u^1, u^2)$ for some $\kappa \neq 0$. Considering now the component

$$\begin{aligned}\mathcal{E}_{11} = (2\partial_2\phi - \kappa\Phi_{22})u_{2'} + 2\partial_{11}\phi - 2\partial_2\phi\,{}^u\Gamma_{11}^2 \\ + 2\phi\partial_2 {}^u\Gamma_{11}^2 + \phi\Phi_{22} + 2\kappa\Phi_{22}{}^u\Gamma_{11}^2 + \kappa\partial_2\Phi_{11} - 2\kappa\partial_1\Phi_{12}\end{aligned}$$

it follows that the conformally Einstein equation reduces to

$$\begin{aligned}\kappa\Phi_{22} &= 2\partial_2\phi, \\ (\phi + 2\kappa\,{}^u\Gamma_{11}^2)\Phi_{22} &= -2\big(\mathrm{Hes}_\phi^D(\partial_{u^1}, \partial_{u^1}) + \phi\rho^D(\partial_{u^1}, \partial_{u^1})\big) + \kappa(2\partial_1\Phi_{12} - \partial_2\Phi_{11}),\end{aligned}$$

from where (ii) is obtained. □

7. Examples

Nilpotent Riemannian extensions with flat base. Let (Σ, D) be a flat torsion-free affine surface. Take local coordinates on Σ so that all Christoffel symbols vanish. Let T be a parallel nilpotent $(1, 1)$-tensor field. Since T is parallel, its components T_i^j are necessarily constant on the given coordinates. Hence one may further specialize the local coordinates (x^1, x^2), by using a linear transformation, so that $T\partial_1 = \partial_2$, $T\partial_2 = 0$ and all the Christoffel symbols ${}^D\Gamma_{ij}^k$ remain identically zero. Now Theorem 3.1 shows that $(T^*\Sigma, g_{D,\Phi,T})$ *is Bach-flat for any symmetric* $(0, 2)$*-tensor field* Φ *on* Σ. Moreover it follows from Theorem 4.3 that $(T^*\Sigma, g_{D,\Phi,T}, f = h \circ \pi)$ *is a steady gradient Ricci soliton for any* $h \in \mathcal{C}^\infty(\Sigma)$ *with* $dh \circ T = 0$ *and any symmetric* $(0, 2)$*-tensor field* Φ *such that* $\Phi_{22}(x^1, x^2) = -h''(x^1)$.

Further note from Remark 4.5 that the steady gradient Ricci soliton

$$(T^*\Sigma, g_{D,\Phi,T}, f = h \circ \pi)$$

satisfies $\mathbb{D} = 0$. Moreover, since $\Phi_{22} = -h''(x^1)$, one has that $(T^*\Sigma, g_{D,\Phi,T})$ is in the conformal class of an Einstein metric (just considering the conformal metric $\bar{g} = \phi^{-2} g_{D,\Phi,T}$ determined by the equation $\phi''(x^1) - \frac{1}{2}\phi(x^1)h''(x^1) = 0$).

Remark 7.1. Set $\Sigma = \mathbb{R}^2$ with usual coordinates (x^1, x^2) and put $T\partial_1 = \partial_2$, $T\partial_2 = 0$. For any smooth function $h(x^1)$ consider the deformation tensor Φ given by $\Phi_{22}(x^1, x^2) = -h''(x^1)$ (the other components being zero). Then, the nonzero Christoffel symbols of $g_{D,\Phi,T}$ are given by

$$\Gamma_{11}^2 = -x_{2'} = -\Gamma_{12'}^{1'}, \quad \Gamma_{11}^{2'} = -h''(x^1)x_{2'}, \quad \Gamma_{12}^{2'} = -\tfrac{1}{2}h^{(3)}(x^1) = -\Gamma_{22}^{1'}.$$

Hence a curve $\gamma(t) = (x^1(t), x^2(t), x_{1'}(t), x_{2'}(t))$ is a geodesic if and only if

$$\begin{gathered}
\ddot{x}^1(t) = 0, \qquad \ddot{x}^2(t) - x_{2'}(t)\,\dot{x}^1(t)^2 = 0,\\
\ddot{x}_{1'}(t) + 2\,x_{2'}(t)\,\dot{x}^1(t)\dot{x}_{2'}(t) + \tfrac{1}{2}\,h^{(3)}(x^1(t))\,\dot{x}^2(t)^2 = 0,\\
\ddot{x}_{2'}(t) - h''(x^1(t))\,x_{2'}(t)\,\dot{x}^1(t)^2 - h^{(3)}(x^1(t))\,\dot{x}^1(t)\,\dot{x}^2(t) = 0.
\end{gathered}$$

Thus $x^1(t) = at + b$ for some $a, b \in \mathbb{R}$ and

$$\begin{gathered}
\ddot{x}^2(t) - a^2\,x_{2'}(t) = 0,\\
\ddot{x}_{2'}(t) - h''(at+b)\,a^2\,x_{2'}(t) - h^{(3)}(at+b)\,a\,\dot{x}^2(t) = 0,\\
\ddot{x}_{1'}(t) + 2a\,x_{2'}(t)\,\dot{x}_{2'}(t) + \tfrac{1}{2}\,h^{(3)}(at+b)\,\dot{x}^2(t)^2 = 0.
\end{gathered}$$

Now the first two equations above are linear and thus $x^2(t)$ and $x_{2'}(t)$ are globally defined. Finally, since $\ddot{x}_{1'}(t) + 2a\,x_{2'}(t)\,\dot{x}_{2'}(t) + \frac{1}{2}\,h^{(3)}(at+b)\,\dot{x}^2(t)^2 = 0$ is also linear on $x_{1'}(t)$, one has that geodesics are globally defined.

Then it follows from Theorem 4.3 that $(T^*\mathbb{R}^2, g_{D,\Phi,T}, f = h\circ\pi)$ *is a geodesically complete steady gradient Ricci soliton, which is conformally Einstein* by Theorem 6.1.

Nilpotent Riemannian extensions with nonrecurrent base. Let $(T^*\Sigma, g_{D,\Phi,T}, f = h\circ\pi)$ be a nontrivial Bach-flat steady gradient Ricci soliton as in Theorem 4.3. Further assume that the Ricci tensor ρ^D is nonsymmetric, i.e., $\rho^D_{\mathrm{sk}} \neq 0$ (equivalently $\partial_2{}^D\Gamma_{11}^1 \neq 0$ as shown in the proof of Proposition 3.3). Then it follows from Theorem 5.1 that $(T^*\Sigma, g_{D,\Phi,T})$ is not half conformally flat.

Theorem 6.1 shows that $(T^*\Sigma, g_{D,\Phi,T})$ is conformally Einstein if and only if there exists a positive $\phi \in \mathcal{C}^\infty(\Sigma)$ with $d\phi\circ T = 0$ such that

$$\phi\,\widehat{\Phi} + 2(\mathrm{Hes}^D_\phi + \phi\,\rho^D_{\mathrm{sym}}) = 0.$$

Hence it follows from Theorem 4.3 that $\mathrm{Hes}^D_h = \frac{2}{\phi}\,\mathrm{Hes}^D_\phi$, which means

$$\Big(2\frac{\phi'}{\phi} - h'\Big){}^D\Gamma_{11}^1 = 2\frac{\phi''}{\phi} - h''.$$

Taking derivatives with respect to x^2 and, since $\partial_2{}^D\Gamma_{11}^1 \neq 0$, the equation above splits into

$$2\frac{\phi'}{\phi} - h' = 0, \quad \text{and} \quad 2\frac{\phi''}{\phi} - h'' = 0,$$

which only admits constant solutions. Summarizing the above one has the following: *Let* (Σ, D, T) *be an affine surface with nonsymmetric Ricci tensor* (*i.e.,* $\rho^D_{\mathrm{sk}} \neq 0$). *Then any Bach-flat gradient Ricci soliton* $(T^*\Sigma, g_{D,\Phi,T}, f = h\circ\pi)$ *is neither half conformally flat nor conformally Einstein.*

Acknowledgements. It is a pleasure to acknowledge useful conversations on this subject with Professor P. Gilkey.

This research was partially supported by projects MTM2013-41335-P, MTM2016-75897-P and EM2014/009 (AEI/FEDER, UE).

References

[Abbena et al. 2013] E. Abbena, S. Garbiero, and S. Salamon, "Bach-flat Lie groups in dimension 4", *C. R. Math. Acad. Sci. Paris* **351**:7-8 (2013), 303–306. MR Zbl

[Bach 1921] R. Bach, "Zur Weylschen Relativitätstheorie und der Weylschen Erweiterung des Krümmungstensorbegriffs", *Math. Z.* **9**:1-2 (1921), 110–135. MR Zbl

[Bakas et al. 2010] I. Bakas, F. Bourliot, D. Lüst, and M. Petropoulos, "Geometric flows in Hořava–Lifshitz gravity", *J. High Energy Phys.* **2010**:4 (2010), art. id. 131. MR Zbl

[Besse 1987] A. L. Besse, *Einstein manifolds*, Ergebnisse der Mathematik (3) **10**, Springer, 1987. MR Zbl

[Brinkmann 1924] H. W. Brinkmann, "Riemann spaces conformal to Einstein spaces", *Math. Ann.* **91**:3-4 (1924), 269–278. MR Zbl

[Brozos-Vázquez and García-Río 2016] M. Brozos-Vázquez and E. García-Río, "Four-dimensional neutral signature self-dual gradient Ricci solitons", *Indiana Univ. Math. J.* **65**:6 (2016), 1921–1943. MR Zbl

[Brozos-Vázquez et al. 2009] M. Brozos-Vázquez, E. García-Río, P. Gilkey, S. Nikčević, and R. Vázquez-Lorenzo, *The geometry of Walker manifolds*, Synthesis Lectures on Mathematics and Statistics **5**, Morgan & Claypool, Williston, VT, 2009. MR Zbl

[Brozos-Vázquez et al. 2013] M. Brozos-Vázquez, E. García-Río, and S. Gavino-Fernández, "Locally conformally flat Lorentzian gradient Ricci solitons", *J. Geom. Anal.* **23**:3 (2013), 1196–1212. MR Zbl

[Brozos-Vázquez et al. 2018] M. Brozos-Vázquez, E. García-Río, and P. Gilkey, "Homogeneous affine surfaces: affine Killing vector fields and gradient Ricci solitons", *J. Math. Soc. Japan* **70**:1 (2018), 25–69.

[Calviño-Louzao et al. 2009] E. Calviño-Louzao, E. García-Río, P. Gilkey, and R. Vázquez-Lorenzo, "The geometry of modified Riemannian extensions", *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **465**:2107 (2009), 2023–2040. MR Zbl

[Cao 2010] H.-D. Cao, "Recent progress on Ricci solitons", pp. 1–38 in *Recent advances in geometric analysis* (Taipei, 2007), edited by Y.-I. Lee et al., Adv. Lect. Math. **11**, Int. Press, Somerville, MA, 2010. MR Zbl arXiv

[Cao and Chen 2012] H.-D. Cao and Q. Chen, "On locally conformally flat gradient steady Ricci solitons", *Trans. Amer. Math. Soc.* **364**:5 (2012), 2377–2391. MR Zbl

[Cao and Chen 2013] H.-D. Cao and Q. Chen, "On Bach-flat gradient shrinking Ricci solitons", *Duke Math. J.* **162**:6 (2013), 1149–1169. MR Zbl

[Cao et al. 2014] H.-D. Cao, G. Catino, Q. Chen, C. Mantegazza, and L. Mazzieri, "Bach-flat gradient steady Ricci solitons", *Calc. Var. Partial Differential Equations* **49**:1-2 (2014), 125–138. MR Zbl

[Chen and He 2013] Q. Chen and C. He, "On Bach flat warped product Einstein manifolds", *Pacific J. Math.* **265**:2 (2013), 313–326. MR Zbl

[Chen and Wang 2015] X. Chen and Y. Wang, "On four-dimensional anti-self-dual gradient Ricci solitons", *J. Geom. Anal.* **25**:2 (2015), 1335–1343. MR Zbl

[Derdzinski 1983] A. Derdzinski, "Self-dual Kähler manifolds and Einstein manifolds of dimension four", *Compositio Math.* **49**:3 (1983), 405–433. MR Zbl

[Derdzinski 2008] A. Derdzinski, "Connections with skew-symmetric Ricci tensor on surfaces", *Results Math.* **52**:3-4 (2008), 223–245. MR Zbl

[Díaz-Ramos et al. 2006] J. C. Díaz-Ramos, E. García-Río, and R. Vázquez-Lorenzo, "Four-dimensional Osserman metrics with nondiagonalizable Jacobi operators", *J. Geom. Anal.* **16**:1 (2006), 39–52. MR Zbl

[Dunajski and Tod 2014] M. Dunajski and K. P. Tod, "Self-dual conformal gravity", *Comm. Math. Phys.* **331**:1 (2014), 351–373. MR Zbl

[Fernández-López and García-Río 2014] M. Fernández-López and E. García-Río, "A note on locally conformally flat gradient Ricci solitons", *Geom. Dedicata* **168** (2014), 1–7. MR Zbl

[Gover and Nagy 2007] A. R. Gover and P.-A. Nagy, "Four-dimensional conformal C-spaces", *Q. J. Math.* **58**:4 (2007), 443–462. MR Zbl

[Hill and Nurowski 2009] C. D. Hill and P. Nurowski, "Intrinsic geometry of oriented congruences in three dimensions", *J. Geom. Phys.* **59**:2 (2009), 133–172. MR Zbl

[Kozameh et al. 1985] C. N. Kozameh, E. T. Newman, and K. P. Tod, "Conformal Einstein spaces", *Gen. Relativity Gravitation* **17**:4 (1985), 343–352. MR Zbl

[Kühnel and Rademacher 2008] W. Kühnel and H.-B. Rademacher, "Conformal transformations of pseudo-Riemannian manifolds", pp. 261–298 in *Recent developments in pseudo-Riemannian geometry*, edited by D. V. Alekseevsky and H. Baum, Eur. Math. Soc., Zürich, 2008. MR Zbl

[Leistner and Nurowski 2010] T. Leistner and P. Nurowski, "Ambient metrics for n-dimensional pp-waves", *Comm. Math. Phys.* **296**:3 (2010), 881–898. MR Zbl

[Munteanu and Sesum 2013] O. Munteanu and N. Sesum, "On gradient Ricci solitons", *J. Geom. Anal.* **23**:2 (2013), 539–561. MR Zbl

[Patterson and Walker 1952] E. M. Patterson and A. G. Walker, "Riemann extensions", *Quart. J. Math., Oxford Ser.* (2) **3** (1952), 19–28. MR Zbl

[Petersen and Wylie 2010] P. Petersen and W. Wylie, "On the classification of gradient Ricci solitons", *Geom. Topol.* **14**:4 (2010), 2277–2300. MR Zbl

[Walker 1950] A. G. Walker, "Canonical form for a Riemannian space with a parallel field of null planes", *Quart. J. Math., Oxford Ser.* (2) **1** (1950), 69–79. MR Zbl

[Wong 1964] Y.-C. Wong, "Two dimensional linear connexions with zero torsion and recurrent curvature", *Monatsh. Math.* **68** (1964), 175–184. MR Zbl

Received July 26, 2016. Revised June 2, 2017.

ESTEBAN CALVIÑO-LOUZAO
CONSELLERÍA DE CULTURA, EDUCACIÓN E ORDENACIÓN UNIVERSITARIA
SANTIAGO DE COMPOSTELA
SPAIN
estebcl@edu.xunta.es

EDUARDO GARCÍA-RÍO
FACULTY OF MATHEMATICS
UNIVERSITY OF SANTIAGO DE COMPOSTELA
SANTIAGO DE COMPOSTELA
SPAIN
eduardo.garcia.rio@usc.es

IXCHEL GUTIÉRREZ-RODRÍGUEZ
FACULTY OF MATHEMATICS
UNIVERSITY OF SANTIAGO DE COMPOSTELA
SANTIAGO DE COMPOSTELA
SPAIN
ixcheldzohara.gutierrez@usc.es

RAMÓN VÁZQUEZ-LORENZO
DEPARTMENT OF MATHEMATICS
IES DE RIBADEO DIONISIO GAMALLO
RIBADEO
SPAIN
ravazlor@edu.xunta.es

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.101

CONTACT STATIONARY LEGENDRIAN SURFACES IN $\mathbb{S}^5$

YONG LUO

Let $(M^5, \alpha, g_\alpha, J)$ be a 5-dimensional Sasakian Einstein manifold with contact 1-form α, associated metric g_α and almost complex structure J, and let L be a contact stationary Legendrian surface in M^5. We will prove that L satisfies the equation

$$-\Delta^\nu H + (K-1)H = 0,$$

where Δ^ν is the normal Laplacian with respect to the metric g on L induced from g_α and K is the Gauss curvature of (L, g).

Using this equation and a new Simons' type inequality for Legendrian surfaces in the standard unit sphere $\mathbb{S}^5$, we prove an integral inequality for contact stationary Legendrian surfaces in $\mathbb{S}^5$. In particular, we prove that if L is a contact stationary Legendrian surface in $\mathbb{S}^5$ and B is the second fundamental form of L, with $S = |B|^2$, $\rho^2 = S - 2H^2$ and

$$0 \le S \le 2,$$

then we have either $\rho^2 = 0$ and L is totally umbilic or $\rho^2 \ne 0$, $S = 2$, $H = 0$ and L is a flat minimal Legendrian torus.

1. Introduction

Let $(M^{2n+1}, \alpha, g_\alpha, J)$ be a $2n+1$ dimensional contact metric manifold with contact structure α, associated metric g_α and almost complex structure J. Assume that (L, g) is an n-dimensional compact Legendrian submanifold of M^{2n+1} with metric g induced from g_α. The volume of L is defined by

$$V(L) = \int_L d\mu, \tag{1-1}$$

where $d\mu$ is the volume form of g. A contact stationary Legendrian submanifold of M^{2n+1} is a Legendrian submanifold of M^{2n+1} which is a stationary point of V with respect to Legendrian deformations. That is we call a Legendrian submanifold

MSC2010: 53B25, 53C42.
Keywords: volume functional, Simons inequality, gap theorem.

$L \subseteq M^{2n+1}$ a contact stationary Legendrian submanifold, if for any Legendrian deformations $L_t \subseteq M^{2n+1}$ with $L_0 = L$ we have

$$\left.\frac{dV(L_t)}{dt}\right|_{t=0} = 0.$$

Remark 1.1. L_t is a Legendrian deformation of $L := L_0$, if L_t is a Legendrian submanifold for every t.

The E–L equation for a contact stationary Legendrian submanifold L is [Iriyeh 2005; Castro et al. 2006]

$$\operatorname{div}_g(JH) = 0, \tag{1-2}$$

where div_g is the divergence with respect to g and H is the mean curvature vector of L in M^{2n+1}.

Remark 1.2. The notion of a contact stationary Legendrian submanifold was first defined by Iriyeh [2005] and Castro et al. [2006] independently, where they used the name of Legendrian minimal Legendrian submanifold and contact minimal Legendrian submanifold, respectively. In this paper we prefer to use the name of contact stationary Legendrian submanifold.

The study of contact stationary Legendrian submanifolds is motivated by the study of Hamiltonian minimal Lagrangian (briefly, HSL) submanifolds, which was first studied by Oh [1990; 1993]. An HSL submanifold in a Kähler manifold is a Lagrangian submanifold which is a stationary point of the Volume functional under Hamiltonian deformations. By [Reckziegel 1988], Legendrian submanifolds in a Sasakian manifold M^{2n+1} can be seem as links of Lagrangian submanifolds in the cone CM^{2n+1}, which is a Kähler manifold with proper metric and complex structure (see Section 2). In fact, a close relation between contact stationary Legendrian submanifolds and HSL submanifolds was found by Iriyeh [2005] and Castro et al. [2006]. Precisely, they independently proved that $C(L)$ is an HSL submanifold in $\mathbb{C}^n (n \geq 2)$ if and only if L is a contact stationary Legendrian submanifold in $\mathbb{S}^{2n-1}$ and L is a contact stationary Legendrian submanifold in $\mathbb{S}^{2n+1} (n \geq 1)$ if and only if $\Pi(L)$ is an HSL submanifold in $\mathbb{CP}^n$, where $\Pi : \mathbb{S}^{2n+1} \to \mathbb{CP}^n$ is the Hopf fibration.

From the definition we see that minimal Legendrian submanifolds are a special kind of contact stationary Legendrian submanifold. Another special kind of contact stationary Legendrian submanifold are Legendrian submanifolds with parallel mean curvature vector fields in the normal bundle. The study of (nonminimal) contact stationary Legendrian submanifolds of $\mathbb{S}^{2n+1}$ is relatively recent endeavor. For $n = 1$, by [Iriyeh 2005], contact stationary Legendrian curves in $\mathbb{S}^3$ are the so called (p, q) curves discovered by Schoen and Wolfson [2001], where p and q are relatively prime integers. For $n = 2$, since a harmonic 1-forms on a 2-sphere must be trivial, contact stationary Legendrian 2-spheres in $\mathbb{S}^5$ must be minimal and so must be the equatorial 2-spheres by Yau's result [1974]. There are a lot of contact stationary doubly periodic surfaces form $\mathbb{R}^2$ to $\mathbb{S}^5$ by lifting Hélein and Romon's examples [2002] and more contact stationary Legendrian surfaces (mainly tori) are constructed in [Mironov 2003; 2008; Iriyeh 2005; Hélein and Romon 2005; Ma 2005; Ma and Schmies 2006; Butscher and Corvino 2012]. And general dimension examples are constructed in [Oh 1993; Mironov 2004; Dong and Han 2007; Dong 2007; Butscher 2009; Joyce et al. 2011; Lee 2012; Chen et al. 2012]. See also [Ono 2005; Hunter and McIntosh 2011; Kajigaya 2013] for other studies of contact stationary Legendrian submanifolds.

In this paper we will study pinching properties of contact stationary Legendrian surfaces in $\mathbb{S}^5$. To do this we first prove an equation satisfied by contact stationary Legendrian surfaces in a Sasakian Einstein manifold, which we hope will be useful in analyzing analytic properties of contact stationary Legendrian surfaces.

Theorem 1.3. *Let L be a contact stationary Legendrian surface in a 5-dimensional Sasakian Einstein manifold $(M^5, \alpha, g_\alpha, J)$, then L satisfies the following equation*:

$$-\Delta^\nu H + (K-1)H = 0, \tag{1-3}$$

where Δ^ν is the normal Laplacian with respect to the metric g on L induced from g_α and K is the Gauss curvature of (L, g).

We recall that the well-known Clifford torus is

$$T_{\mathrm{Clif}} = \mathbb{S}^1\Big(\frac{1}{\sqrt{2}}\Big) \times \mathbb{S}^1\Big(\frac{1}{\sqrt{2}}\Big) \subseteq \mathbb{S}^5. \tag{1-4}$$

In the theory of minimal surfaces, the following Simons' integral inequality and pinching theorem due to Simons [1968], Lawson [1969] and Chern et al. [1970] are well known.

Theorem 1.4. *Let M be a compact minimal surface in a unit sphere $\mathbb{S}^3$ and B be the second fundamental form of M in $\mathbb{S}^3$. Set $S = |B|^2$, then we have*

$$\int_M S(2-S)\, d\mu \le 0.$$

In particular, *if*

$$0 \le S \le 2,$$

then either $S = 0$ and M is totally geodesic, *or $S = 2$ and M is the Clifford torus T_{Clif}*, *which is defined by* (1-4).

The above integral inequality was proved by Simons [1968] in his celebrated paper and the classification result was given by Chern et al. [1970] and Lawson [1969], independently.

For minimal surfaces in a sphere with higher codimension, a corresponding integral inequality was proved by Benko et al. [1979] and Kozlowski and Simon [1984]. In order to state their result, we first record an example.

Example. The Veronese surface is a minimal surface in $\mathbb{S}^4 \subseteq \mathbb{R}^5$ defined by

$$u : \mathbb{S}^2(\sqrt{3}) \subseteq \mathbb{R}^3 \to \mathbb{S}^4(1) \subseteq \mathbb{R}^5$$
$$(x, y, z) \mapsto (u_1, u_2, u_3, u_4, u_5)$$

where

$$u_1 = \frac{yz}{\sqrt{3}}, \quad u_2 = \frac{xz}{\sqrt{3}}, \quad u_3 = \frac{xy}{\sqrt{3}}, \quad u_4 = \frac{x^2 - y^2}{2\sqrt{3}}, \quad u_5 = \frac{x^2 + y^2 - 2z^2}{6}.$$

Here, u defines an isometric immersion of $\mathbb{S}^2(\sqrt{3})$ into $\mathbb{S}^4(1)$, and it maps two points (x, y, z) and $(-x, -y, -z)$ of $\mathbb{S}^2(\sqrt{3})$ into the same point of $\mathbb{S}^4(1)$, and so it imbeds the real projective plane into $\mathbb{S}^4(1)$.

We have

Theorem 1.5 [Benko et al. 1979]. *Let M be a minimal surface in an n-dimensional sphere $\mathbb{S}^n$, then*

$$\int_M S\left(2 - \tfrac{3}{2}S\right) d\mu \leq 0. \tag{1-5}$$

In particular, if

$$0 \leq S \leq \tfrac{4}{3},$$

then either $S = 0$ and M is totally geodesic, or $S = \frac{4}{3}$, $n = 4$ and M is the Veronese surface.

The above classification for minimal surfaces in a sphere with $S = \frac{4}{3}$ was also shown by Chern et al. [1970].

We see that the (first) pinching constant for minimal surfaces in $\mathbb{S}^3$ is 2, but it is $\frac{4}{3}$ for minimal surfaces of higher codimension. This is an interesting phenomenon and we think it is due to the complexity of the normal bundle, because for minimal Legendrian surfaces in $\mathbb{S}^5$, the (first) pinching constant is also 2.

Theorem 1.6 [Yamaguchi et al. 1976]. *If M is a minimal Legendrian surface of the unit sphere $\mathbb{S}^5$ and $0 \leq S \leq 2$, then S is identically 0 or 2.*

Remark 1.7. For higher dimensional case of this theorem we refer to [Dillen and Vrancken 1990].

All of these results are based on calculating the Laplacian of S and then getting Simons' type equalities or inequalities, a powerful method which was originated by Simons [1968]. The minimal condition is used to cancel some terms in the resulting calculation and to some extent it is important. In this note we prove a Simons' type inequality (Lemma 3.8) for Legendrian surfaces in $\mathbb{S}^5$, without minimal condition. By using (1-3) and this Simons' type inequality we get

Theorem 1.8. *Let $L : \Sigma \to \mathbb{S}^5$ be a contact stationary Legendrian surface, where $\mathbb{S}^5$ is the unit sphere with standard contact structure and metric (as given at the end of Section 2). Then we have*

$$\int_L \rho^2 \big(3 - \tfrac{3}{2} S + 2H^2\big)\, d\mu \le 0,$$

where $\rho^2 := S - 2H^2$. In particular, if

$$0 \le S \le 2,$$

then either $\rho^2 = 0$ and L is totally umbilic, or $\rho^2 \ne 0$, $S = 2$, $H = 0$ and L is a flat minimal Legendrian torus.

Remark 1.9. Because minimal Legendrian surfaces are contact stationary Legendrian surfaces and satisfy $\rho^2 = S$, and because totally umbilic minimal surfaces are totally geodesic, we see that Theorem 1.6 is a corollary of Theorem 1.8.

Integral inequality and gap phenomenon for submanifolds satisfying a fourth order quasielliptic nonlinear equation was first studied by Li [2001; 2002a; 2002b] who proved several gap theorems for Willmore submanifolds in a sphere. These results are partial motivations of our paper.

We end this introduction by recalling a classification theorem of flat minimal Legendrian tori in $\mathbb{S}^5$. For a constant θ let T_θ be the 2-torus in $\mathbb{S}^5$ defined by

$$T_\theta = \Big\{(z_1, z_2, z_3) \in \mathbb{C}^3 : |z_i| = \tfrac{1}{3}, i = 1, 2, 3 \text{ and } \sum_i \arg z_i = \theta\Big\}.$$

T_θ is called the generalized Clifford torus and it is a flat minimal Legendrian torus in $\mathbb{S}^5$. Its projection under the Hopf map $\pi : \mathbb{S}^5 \to \mathbb{CP}^2$ is a flat minimal Lagrangian torus, which is also called a generalized Clifford torus. It is proved in [Ludden et al. 1975] that a flat minimal Lagrangian torus in $\mathbb{CP}^2$ must be $\mathbb{S}^1 \times \mathbb{S}^1$. By the correspondence of minimal Lagrangian surfaces in $\mathbb{CP}^2$ and minimal Legendrian surfaces in $\mathbb{S}^5$ (see [Reckziegel 1988]), we see that a flat minimal Legendrian torus in $\mathbb{S}^5$ must be a generalized Clifford torus. For more details we refer to [Haskins 2004, p. 853].

The rest of this paper is organized as follows: In Section 2 we collect some basic material from Sasakian geometry, which will be used in the next section. In Section 3 we prove our main results, Theorems 1.3 and 1.8.

2. Preliminaries on contact geometry

In this section we recall some basic material from contact geometry. For more information we refer to [Blair 2002].

Contact manifolds.

Definition 2.1. A contact manifold M is an odd dimensional manifold with a one form α such that $\alpha \wedge (d\alpha)^n \neq 0$, where $\dim M = 2n+1$.

Assume now that (M, α) is a given contact manifold of dimension $2n+1$. Then α defines a $2n$-dimensional vector bundle over M, where the fiber at each point $p \in M$ is given by

$$\xi_p = \operatorname{Ker} \alpha_p.$$

Since $\alpha \wedge (d\alpha)^n$ defines a volume form on M, we see that

$$\omega := d\alpha$$

is a closed nondegenerate 2-form on $\xi \oplus \xi$ and hence it defines a symplectic product on ξ such that $(\xi, \omega|_{\xi\oplus\xi})$ becomes a symplectic vector bundle. A consequence of this fact is that there exists an almost complex bundle structure

$$\tilde{J} : \xi \to \xi$$

compatible with $d\alpha$, i.e., a bundle endomorphism satisfying

(1) $\tilde{J}^2 = -\operatorname{id}_\xi$,
(2) $d\alpha(\tilde{J}X, \tilde{J}Y) = d\alpha(X, Y)$ for all $X, Y \in \xi$,
(3) $d\alpha(X, \tilde{J}X) > 0$ for $X \in \xi \setminus 0$.

Since M is an odd dimensional manifold, ω must be degenerate on TM, and so we obtains a line bundle η over M with fibers

$$\eta_p := \{V \in T_pM | \omega(V, W) = 0, \forall W \in \xi_p\}.$$

Definition 2.2. The Reeb vector field $\boldsymbol{R}$ is the section of η such that $\alpha(\boldsymbol{R}) = 1$.

Thus α defines a splitting of TM into a line bundle η with the canonical section $\boldsymbol{R}$ and a symplectic vector bundle $(\xi, \omega|\xi \oplus \xi)$. We denote the projection along η by π, i.e.,

$$\pi : TM \to \xi, \quad \pi(V) := V - \alpha(V)\boldsymbol{R}.$$

Using this projection we extend the almost complex structure $\tilde{J}$ to a section $J \in \Gamma(T^*M \otimes TM)$ by setting

$$J(V) = \tilde{J}(\pi(V)),$$

for $V \in TM$.

We call J an almost complex structure of the contact manifold M.

Definition 2.3. Let (M, α) be a contact manifold, a submanifold L of (M, α) is called an isotropic submanifold if $T_xL \subseteq \xi_x$ for all $x \in L$.

For algebraic reasons the dimension of an isotropic submanifold of a $2n+1$ dimensional contact manifold can not be bigger than n.

Definition 2.4. An isotropic submanifold $L \subseteq (M, \alpha)$ of maximal possible dimension n is called a Legendrian submanifold.

Sasakian manifolds. Let (M, α) be a contact manifold, with the almost complex structure J and Reeb field $\boldsymbol{R}$. A Riemannian metric g_α defined on M is said to be associated, if it satisfies the following three conditions:

(1) $g_\alpha(\boldsymbol{R}, \boldsymbol{R}) = 1$.

(2) $g_\alpha(V, \boldsymbol{R}) = 0, \forall V \in \xi$.

(3) $\omega(V, JW) = g_\alpha(V, W), \forall V, W \in \xi$.

We should mention here that on any contact manifold there exists an associated metric on it, because we can construct one in the following way. We introduce a bilinear form b by

$$b(V, W) := \omega(V, JW),$$

then the tensor

$$g := b + \alpha \otimes \alpha$$

defines an associated metric on M.

Sasakian manifolds are the odd dimensional analogue of Kähler manifolds. They are defined as follows.

Definition 2.5. A contact manifold (M, α) with an associated metric g_α is called Sasakian, if the cone CM equipped with the following extended metric $\bar{g}$

$$(CM, \bar{g}) = (\mathbb{R}_+ \times M, dr^2 + r^2 g_\alpha) \tag{2-1}$$

is Kähler with respect to the following canonical almost complex structure J on $TCM = \mathbb{R} \oplus \langle \boldsymbol{R} \rangle \oplus \xi$:

$$J(r\partial r) = \boldsymbol{R},\ J(\boldsymbol{R}) = -r\partial r.$$

Furthermore if g_α is Einstein, M is called a Sasakian Einstein manifold.

We record several lemmas which are well known in Sasakian geometry. These lemmas will be used in the next section.

Lemma 2.6. *Let (M, α, g_α, J) be a Sasakian manifold. Then*

$$\overline{\nabla}_X \boldsymbol{R} = -JX, \tag{2-2}$$

and

$$(\overline{\nabla}_X J)(Y) = g(X, Y)\boldsymbol{R} - \alpha(Y)X, \tag{2-3}$$

for $X, Y \in TM$, where $\overline{\nabla}$ is the Levi–Civita connection on (M, g_α).

Lemma 2.7. *Let L be a Legendrian submanifold in a Sasakian Einstein manifold (M, α, g_α, J), then the mean curvature form $\omega(H, \cdot)|_L$ defines a closed one form on L.*

For a proof of this lemma we refer to [Lê 2004, Proposition A.2] or [Smoczyk 2003, Lemma 2.8]. In fact they proved this result under the weaker assumption that (M, α, g_α, J) is a weakly Sasakian Einstein manifold, where weakly Einstein means that g_α is Einstein only when restricted to the contact hyperplane $\operatorname{Ker}\alpha$.

Lemma 2.8. *Let L be a Legendrian submanifold in a Sasakian manifold (M,α,g_α,J) and B be the second fundamental form of L in M. Then we have*

$$g_\alpha(B(X, Y), \boldsymbol{R}) = 0, \tag{2-4}$$

for any $X, Y \in TL$.

Proof. For any $X, Y \in TL$,

$$\begin{aligned}\langle B(X, Y), \boldsymbol{R}\rangle &= \langle \overline{\nabla}_X Y, \boldsymbol{R}\rangle\\ &= -\langle Y, \overline{\nabla}_X \boldsymbol{R}\rangle\\ &= \langle Y, JX\rangle\\ &= \omega(X, Y)\\ &= d\alpha(X, Y)\\ &= 0,\end{aligned}$$

where in the third equality we used (2-2). □

In particular this lemma implies that the mean curvature H of L is orthogonal to the Reeb field $\boldsymbol{R}$.

Lemma 2.9. *For any $Y, Z \in \operatorname{Ker}\alpha$, we have*

$$g_\alpha(\overline{\nabla}_X(JY), Z) = g_\alpha(J\overline{\nabla}_X Y, Z). \tag{2-5}$$

Proof. Note that

$$(\overline{\nabla}_X J)Y = \overline{\nabla}_X(JY) - J\overline{\nabla}_X Y.$$

Therefore by using (2-3) we have

$$\begin{aligned}\langle \overline{\nabla}_X(JY), Z\rangle &= \langle (\overline{\nabla}_X J)Y, Z\rangle + \langle J\overline{\nabla}_X Y, Z\rangle\\ &= \langle J\overline{\nabla}_X Y, Z\rangle,\end{aligned}$$

for any $Y, Z \in \operatorname{Ker}\alpha$. □

A canonical example of Sasakian Einstein manifolds is the standard odd dimensional sphere $\mathbb{S}^{2n+1}$.

The standard sphere $\mathbb{S}^{2n+1}$. Let $\mathbb{C}^n = \mathbb{R}^{2n+2}$ be the Euclidean space with coordinates $(x_1, \dots, x_{n+1}, y_1, \dots, y_{n+1})$ and $\mathbb{S}^{2n+1}$ be the standard unit sphere in $\mathbb{R}^{2n+2}$. Define

$$\alpha_0 = \frac{1}{2}\sum_{j+1}^{n+1}(x_j dy_j - y_j dx_j),$$

then

$$\alpha := \alpha_0|_{\mathbb{S}^{2n+1}}$$

defines a contact one form on $\mathbb{S}^{2n+1}$. Assume that g_0 is the standard metric on $\mathbb{R}^{2n+2}$ and J_0 is the standard complex structure of $\mathbb{C}^n$. We define

$$g_\alpha = g_0|_{\mathbb{S}^{2n+1}} \quad \text{and} \quad J = J_0|_{\mathbb{S}^{2n+1}},$$

then $(\mathbb{S}^{2n+1}, \alpha, g_\alpha, J)$ is a Sasakian Einstein manifold with associated metric g_α. Its contact hyperplane is characterized by

$$\operatorname{Ker}\alpha_x = \{Y \in T_x\mathbb{S}^{2n+1} | \langle Y, Jx\rangle = 0\}.$$

3. Proof of the theorems

Several lemmas. In this part we assume that (M, α, g_α, J) is a Sasakian manifold. We show several lemmas which are analogous to results in Kähler geometry.

The first lemma shows $\omega = d\alpha$ when restricted to the contact hyperplane $\operatorname{Ker}\alpha$ behaves as the Kähler form on a Kähler manifold.

Lemma 3.1. *Let $X, Y, Z \in \operatorname{Ker}\alpha$, then*

$$\overline{\nabla}_X\omega(Y, Z) = 0, \tag{3-1}$$

where $\overline{\nabla}$ is the derivative with respect to g_α.

Proof.

$$\begin{aligned}\overline{\nabla}_X\omega(Y, Z) &= X(\omega(Y, Z)) - \omega(\overline{\nabla}_X Y, Z) - \omega(Y, \overline{\nabla}_X Z)\\ &= -Xg_\alpha(Y, JZ) - \omega(\overline{\nabla}_X Y, Z) - \omega(Y, \overline{\nabla}_X Z)\\ &= -g_\alpha(\overline{\nabla}_X Y, JZ) - g_\alpha(Y, \overline{\nabla}_X JZ) + g_\alpha(\overline{\nabla}_X Y, JZ) + g_\alpha(Y, J\overline{\nabla}_X Z)\\ &= 0,\end{aligned}$$

where in the third equality we used $g_\alpha(Y, \overline{\nabla}_X JZ) = g_\alpha(Y, J\overline{\nabla}_X Z)$, which is a direct corollary of (2-3). □

Now let L be a Legendrian submanifold of M. We have a natural identification of $NL \cap \operatorname{Ker}\alpha$ with T^*L, where NL is the normal bundle of L and T^*L is the cotangent bundle.

Definition 3.2. $\tilde{\omega} : NL \cap \operatorname{Ker}\alpha \to T^*L$ is the bundle isomorphism defined by

$$\tilde{\omega}_p(v_p) = (v_p \rfloor \omega_p)|_{T_pL},$$

where $p \in L$ and $v_p \in (NL \cap \operatorname{Ker}\alpha)_p$.

Recall that $\omega(\boldsymbol{R}) = 0$ and $g_\alpha(V, W) = \omega(V, JW)$ for any $V, W \in \xi$, hence $\tilde{\omega}$ defines an isomorphism.

We have

Lemma 3.3. *Let $V \in \Gamma(NL \cap \operatorname{Ker}\alpha)$. Then*

$$\tilde{\omega}(\Delta^\nu V - \langle \Delta^\nu V, \boldsymbol{R}\rangle \boldsymbol{R} + V) = \Delta(\tilde{\omega}(V)), \quad \textit{i.e.},$$

$$(\Delta^\nu V + V)\rfloor\omega = \Delta(V \rfloor \omega), \tag{3-2}$$

where Δ is the Laplace–Beltrami operator on (L, g).

Remark 3.4. This kind of lemma in the context of symplectic geometry was proved by Oh [1990, Lemma 3.3]. Our proof follows his argument with only slight modifications.

Proof. We first show that

$$\nabla_X(\tilde{\omega}(V)) = \tilde{\omega}\big(\nabla^\nu_X V - \langle \nabla^\nu_X V, \boldsymbol{R}\rangle \boldsymbol{R}\big) \tag{3-3}$$

for any $X \in TL$. Equality (3-3) is equivalent to

$$\nabla_X(\tilde{\omega}(V))(Y) = \tilde{\omega}\big(\nabla^\nu_X V - \langle \nabla^\nu_X V, \boldsymbol{R}\rangle \boldsymbol{R}\big)(Y) \tag{3-4}$$

for any $Y \in TL$.

$$\begin{aligned}
\nabla_X(\tilde{\omega}(V))(Y) &= \nabla_X(\tilde{\omega}(V)(Y)) - \tilde{\omega}(V)(\nabla_X Y) \\
&= \overline{\nabla}_X(\omega(V, Y)) - \tilde{\omega}(V)(\nabla_X Y) \\
&= \omega(\nabla^\nu_X V, Y) + \omega(V, \nabla_X Y) - \omega(V, \nabla_X Y) \\
&= \omega(\nabla^\nu_X V, Y) \\
&= \tilde{\omega}\big(\nabla^\nu_X V - \langle \nabla^\nu_X V, \boldsymbol{R}\rangle \boldsymbol{R}\big)(Y),
\end{aligned}$$

where in the third equality we used $\overline{\nabla}_X \omega = 0$, when restricted to $\operatorname{Ker}\alpha$, which is proved in Lemma 3.1.

Let $p \in L$ and choose an orthonormal frame $\{E_1, \dots, E_n\}$ on TL such that $\nabla_{E_i} E_j(p) = 0$, then the general Laplacian Δ can be written as

$$\Delta\psi(p) = \sum_{i=1}^{n} \nabla_{E_i} \nabla_{E_i} \psi(p),$$

where ψ is a tensor on L. Therefore

$$
\begin{aligned}
(\tilde\omega^{-1}\circ\Delta\cdot\tilde\omega(V))(p)&\\
&=\Big(\tilde\omega^{-1}\circ\sum_{i=1}^n\nabla_{E_i}\nabla_{E_i}\tilde\omega(V)\Big)(p)\\
&=\sum_{i=1}^n(\tilde\omega^{-1}\nabla_{E_i}\tilde\omega\cdot\tilde\omega^{-1}\nabla_{E_i}\tilde\omega(V))(p)\\
&=\sum_{i=1}^n\big(\tilde\omega^{-1}\nabla_{E_i}\tilde\omega\big(\nabla^\nu_{E_i}V-\langle\nabla^\nu_{E_i}V,\boldsymbol{R}\rangle\boldsymbol{R}\big)\big)(p)\\
&=\sum_{i=1}^n\nabla^\nu_{E_i}\big(\nabla^\nu_{E_i}V-\langle\nabla^\nu_{E_i}V,\boldsymbol{R}\rangle\boldsymbol{R}\big)-\big\langle\nabla^\nu_{E_i}\big(\nabla^\nu_{E_i}V-\langle\nabla^\nu_{E_i}V,\boldsymbol{R}\rangle\boldsymbol{R}\big),\boldsymbol{R}\big\rangle\boldsymbol{R}\\
&=\Delta^\nu V-\langle\Delta^\nu V,\boldsymbol{R}\rangle\boldsymbol{R}+V,
\end{aligned}
$$

where in the third and fourth equalities we used (3-3) and in the last equality we used equality (2-2). □

Proof of Theorem 1.3. We see that for any function s defined on L,

$$
\begin{aligned}
0=\int_L s\operatorname{div}JH\,d\mu\ &=\int_L g(JH,\nabla s)\,d\mu\\
=\int_L\omega(H,\nabla s)\,d\mu&=\int_L\langle\omega\rfloor H,\omega\rfloor\nabla s\rangle\,d\mu\\
=\int_L\langle\omega\rfloor H,\ ds\rangle\ &=\int_L\delta(\omega\rfloor H)s\,d\mu.
\end{aligned}
$$

Therefore the E–L equation for L is equivalent to

$$\delta(\omega\rfloor H)=0, \tag{3-5}$$

where δ is the adjoint operator of d on L.

By Lemma 2.7 we see that L satisfies

$$\Delta_h(\omega\rfloor H)=0, \tag{3-6}$$

where $\Delta_h:=\delta d+d\delta$ is the Hodge–Laplace operator. That is the mean curvature form of L is a harmonic one form.

To proceed on, we need the following Weitzenböck formula

Lemma 3.5. *Let M be an n dimensional oriented Riemannian manifold. If $\{V_i\}$ is a local orthonormal frame field and $\{\omega^i\}$ is its dual coframe field, then*

$$\Delta_h=-\sum_i D^2_{V_iV_i}+\sum_{ij}\omega^i\wedge i(V_j)R_{V_iV_j},$$

where $D^2_{XY} \equiv D_X D_Y - D_{D_X Y}$ *represents the covariant derivatives*, $\Delta_d = d\delta + \delta d$ *is the Hodge–Laplace and* $R_{XY} = -D_X D_Y + D_Y D_X + D_{[X,Y]}$ *is the curvature tensor.*

Remark 3.6. For a detailed discussion on the Weitzenböck formula we refer to Wu [1988].

Using the Weitzenböck formula we have

$$-\Delta(\omega \rfloor H) + \sum_{ij} \omega^i \wedge i(V_j) R_{V_i V_j}(\omega \rfloor H) = 0, \tag{3-7}$$

where $\{V_i\}$ is a local orthogonal frame field and $\{\omega^i\}$ is its dual coframe field on L.

Denote $\omega \rfloor H$ by $\theta_H = \sum_k \theta_k \omega^k$, we have

$$\begin{aligned}
\sum_{ij} \omega^i \wedge i(V_j) R_{V_i V_j} \theta_H &= \sum_{ij} R_{V_i V_j} \theta_H(V_j) \omega^i \\
&= \sum_{ijk} R_{V_i V_j} \omega^k(V_j) \theta_k \omega^i \\
&= -\sum_{ijk} \omega^k(R_{V_i V_j} V_j) \theta_k \omega^i \\
&= -\sum_{ijk} \langle R_{V_i V_j} V_j, V_k \rangle \theta_k \omega^i \\
&= -\sum_{ij} \langle R_{V_i V_j} V_j, V_i \rangle \theta_i \omega^i \\
&= K\theta_H.
\end{aligned}$$

That is

$$\sum_{ij} \omega^i \wedge i(V_j) R_{V_i V_j}(\omega \rfloor H) = K\omega \rfloor H. \tag{3-8}$$

Recall that $H \in NL \cap \operatorname{Ker}\alpha$, using (3-2) on H we get

$$\Delta(\omega \rfloor H) = (\Delta^\nu H + H) \rfloor \omega. \tag{3-9}$$

Combining (3-7)–(3-9), we have

$$0 = -\Delta^\nu H \rfloor \omega - H + K\omega \rfloor H = (-\Delta^\nu H + (K-1)H) \rfloor \omega,$$

which implies that

$$-\Delta^\nu H + (K-1)H = f\boldsymbol{R} \tag{3-10}$$

for some function f on L.

The next lemma is one of our key observations which states that a Legendrian submanifold in a Sasakian manifold is contact stationary if and only if $\langle \Delta^\nu H, \boldsymbol{R} \rangle = 0$.

Lemma 3.7. *Let $L \subseteq (M^{2n+1}, \alpha, g_\alpha, J)$ be a contact stationary Legendrian submanifold. Then $\Delta^\nu H$ is orthogonal to $\boldsymbol{R}$.*

Proof. For any point $p \in L$, we choose a local orthonormal frame $\{E_i : i = 1, \dots, n\}$ of L such that $\nabla_{E_i} E_j(p) = 0$. We have at p (in the following computation we adopt the Einstein summation convention)

$$\begin{aligned}
\langle \Delta^\nu H, \boldsymbol{R}\rangle &= \langle \nabla^\nu_{E_i} \nabla^\nu_{E_i} H, \boldsymbol{R}\rangle \\
&= E_i \langle \nabla^\nu_{E_i} H, \boldsymbol{R}\rangle - \langle \nabla^\nu_{E_i} H, \overline{\nabla}_{E_i} \boldsymbol{R}\rangle \\
&= E_i \langle \nabla^\nu_{E_i} H, \boldsymbol{R}\rangle + \langle \nabla^\nu_{E_i} H, JE_i\rangle \\
&= E_i (E_i \langle H, \boldsymbol{R}\rangle - \langle H, \overline{\nabla}_{E_i} \boldsymbol{R}\rangle) + \langle \nabla^\nu_{E_i} H, JE_i\rangle \\
&= E_i \langle H, JE_i\rangle + \langle \nabla^\nu_{E_i} H, JE_i\rangle \\
&= 2\langle \nabla^\nu_{E_i} H, JE_i\rangle + \langle H, \overline{\nabla}_{E_i} JE_i\rangle \\
&= 2\langle \nabla^\nu_{E_i} H, JE_i\rangle + \langle H, J\overline{\nabla}_{E_i} E_i\rangle \\
&= 2\langle \nabla^\nu_{E_i} H, JE_i\rangle \\
&= 2\langle \overline{\nabla}_{E_i} H, JE_i\rangle \\
&= -2\langle J\overline{\nabla}_{E_i} H, E_i\rangle \\
&= -2\langle \overline{\nabla}_{E_i} JH, E_i\rangle \\
&= -2\langle \nabla_{E_i} JH, E_i\rangle \\
&= -2\,\mathrm{div}_g(JH) \\
&= 0.
\end{aligned}$$

Note that in this computation we used Equation (2-3) and Lemmas 2.8 and 2.9 several times and the last equality holds because L is contact stationary. $\square$

Therefore we have

$$(-\Delta^\nu H + (K-1)H) \perp \boldsymbol{R}$$

by this lemma and Lemma 2.8, which shows $f \equiv 0$, i.e.,

$$-\Delta^\nu H + (K-1)H = 0,$$

and we are done with the proof of Theorem 1.3.

Proof of Theorem 1.8. Let L be a Legendrian surface in $\mathbb{S}^5$ with the induced metric g. Let $\{e_1, e_2\}$ be an orthogonal frame on L such that $\{e_1, e_2, Je_1, Je_2, \boldsymbol{R}\}$ is an orthonormal frame on $\mathbb{S}^5$.

In the following we use indices i, j, k, l, s, t, m and β and γ such that

$$1 \le i, j, k, l, s, t, m \le 2, \quad 1 \le \beta, \gamma \le 3, \quad \gamma^* = \gamma + 2 \quad \text{and} \quad \beta^* = \beta + 2.$$

Let B be the second fundamental form of L in $\mathbb{S}^5$ and define

$$h^k_{ij} = g_\alpha(B(e_i, e_j), Je_k), \tag{3-11}$$

$$h^3_{ij} = g_\alpha(B(e_i, e_j), \boldsymbol{R}). \tag{3-12}$$

Then

$$h^k_{ij} = h^j_{ik} = h^i_{kj}, \tag{3-13}$$

$$h^3_{ij} = 0. \tag{3-14}$$

The Gauss equations and Ricci equations are

$$R_{ijkl} = (\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}) + \sum_s (h^s_{ik}h^s_{jl} - h^s_{il}h^s_{jk}) \tag{3-15}$$

$$R_{ik} = \delta_{ik} + 2\sum_s H^s h^s_{ik} - \sum_{s,j} h^s_{ij}h^s_{jk}, \tag{3-16}$$

$$2K = 2 + 4H^2 - S, \tag{3-17}$$

$$\begin{aligned} R_{3412} &= \sum_i (h^1_{i1}h^2_{i2} - h^1_{i2}h^2_{i1}) \\ &= \det h^1 + \det h^2, \end{aligned} \tag{3-18}$$

where h^1 and h^2 are the second fundamental forms with respect to the directions Je_1 and Je_2.

In addition we have the following Codazzi equations and Ricci identities

$$h^\beta_{ijk} = h^\beta_{ikj}, \tag{3-19}$$

$$h^\beta_{ijkl} - h^\beta_{ijlk} = \sum_m h^\beta_{mj} R_{mikl} + \sum_m h^\beta_{mi} R_{mjkl} + \sum_\gamma h^\gamma_{ij} R_{\gamma^*\beta^* kl}. \tag{3-20}$$

Using these equations, we can get the following Simons' type inequality:

Lemma 3.8. *Let L be a Legendrian surface in $\mathbb{S}^5$. Then we have*

$$\begin{aligned} \frac{1}{2}\Delta \sum_{i,j,\beta} (h^\beta_{ij})^2 \geq |\nabla^T h|^2 - 2|\nabla^T H|^2 - 2|\nabla^\nu H|^2 + \sum_{i,j,k,\beta} (h^\beta_{ij}h^\beta_{kki})_j \\ + S - 2H^2 + 2(1+H^2)\rho^2 - \rho^4 - \frac{1}{2}S^2, \end{aligned} \tag{3-21}$$

where $|\nabla^T h|^2 = \sum_{i,j,k,s}(h^s_{ijk})^2$ *and* $|\nabla^T H|^2 = \sum_{i,s}(H^s_i)^2$.

Proof. Using equations (3-15)–(3-20), we have

$$
\begin{aligned}
(3\text{-}22)\quad \frac{1}{2}\Delta\sum_{i,j,\beta}(h^\beta_{ij})^2
&= \sum_{i,j,k,\beta}(h^\beta_{ijk})^2+\sum_{i,j,k,\beta}h^\beta_{ij}h^\beta_{kijk}\\
&=|\nabla h|^2-4|\nabla^\nu H|^2+\sum_{i,j,k,\beta}(h^\beta_{ij}h^\beta_{kki})_j+\sum_{i,j,l,k,\beta}h^\beta_{ij}(h^\beta_{lk}R_{lijk}+h^\beta_{il}R_{lj})\\
&\qquad+\sum_{i,j,k,\beta,\gamma}h^\beta_{ij}h^\gamma_{ki}R_{\gamma^*\beta^*jk}\\
&=|\nabla h|^2-4|\nabla^\nu H|^2+\sum_{i,j,k,s}(h^s_{ij}h^s_{kki})_j+2K\rho^2-2(\det h^1+\det h^2)^2\\
&\geq|\nabla h|^2-4|\nabla^\nu H|^2+\sum_{i,j,k,\beta}(h^\beta_{ij}h^\beta_{kki})_j+2(1+H^2)\rho^2-\rho^4-\tfrac{1}{2}S^2,
\end{aligned}
$$

where $\rho^2:=S-2H^2$ and in the above calculations we used the identities

$$
\begin{aligned}
\sum_{i,j,k,l,\beta}h^\beta_{ij}(h^\beta_{lk}R_{lijk}+h^\beta_{il}R_{lj})&=2K\rho^2,\\
\sum_{i,j,k,\beta,\gamma}h^\beta_{ij}h^\gamma_{ki}R_{\gamma^*\beta^*jk}&=-2(\det h^1+\det h^2)^2,
\end{aligned}
$$

where in the first equality we used $R_{lijk}=K(\delta_{lj}\delta_{ik}-\delta_{lk}\delta_{ij})$ and $R_{lj}=K\delta_{lj}$ in a proper coordinate, because L is a surface.

Note that

$$
\begin{aligned}
(3\text{-}23)\quad |\nabla h|^2=\sum_{i,j,k,\beta}(h^\beta_{ijk})^2&=|\nabla^T h|^2+\sum_{i,j,k}(h^3_{ijk})^2\\
&=|\nabla^T h|^2+\sum_{i,j,k}(h^k_{ij})^2=|\nabla^T h|^2+S,
\end{aligned}
$$

where in the third equality we used

$$
\begin{aligned}
h^3_{ijk}&=\langle\overline{\nabla}_{e_k}B(e_i,e_j),\boldsymbol{R}\rangle=-\langle B(e_i,e_j),\overline{\nabla}_{e_k}\boldsymbol{R}\rangle\\
&=\langle B(e_i,e_j),Je_k\rangle=h^k_{ij}.
\end{aligned}
$$

Similarly we have

$$
(3\text{-}24)\qquad |\nabla^\nu H|^2=|\nabla^T H|^2+H^2.
$$

Combing (3-22), (3-23) and (3-24) we get (3-21). □

Now we prove an integral equality for L, by using (1-3).

Lemma 3.9. *Let $L : \Sigma \to \mathbb{S}^5$ be a contact stationary Legendrian surface, where $\mathbb{S}^5$ is the unit sphere with standard contact structure and metric. Then*

$$\int_L |\nabla^\nu H|^2 \, d\mu = -\int_L (K-1) H^2 \, d\mu, \tag{3-25}$$

where $|\nabla^\nu H|^2 = \sum_{\beta,i} (H_i^\beta)^2$.

Proof. By using (1-3) we have

$$\begin{aligned} |\nabla^\nu H|^2 &= \sum_{\beta,i} (H_i^\beta)^2 \\ &= \sum_{\beta,i} (H_i^\beta H^\beta)_i - \sum_\beta H^\beta \Delta^\nu H^\beta \\ &= \sum_{\beta,i} (H_i^\beta H^\beta)_i - (K-1) H^2. \end{aligned} \tag{3-26}$$

We get (3-25) by integrating over (3-26). □

Integrating over (3-21) and using $|\nabla^T h|^2 \geq 3|\nabla^T H|^2$ (see Lemma A.1) we get

$$\begin{aligned} 0 &\geq \int_L \left[(|\nabla^T h|^2 - 2|\nabla^T H|^2) - 2|\nabla^\nu H|^2 + S - 2H^2 + 2(1+H^2)\rho^2 - \rho^4 - \tfrac{1}{2} S^2 \right] d\mu \\ &\geq \int_L \left[-2|\nabla^\nu H|^2 + S - 2H^2 + 2(1+H^2)\rho^2 - \rho^4 - \tfrac{1}{2} S^2 \right] d\mu \\ &= \int_L (2-\rho^2)\rho^2 d\mu + \int_L 2H^2\rho^2 + 2(K-1)H^2 - 2H^2 + S - \tfrac{1}{2} S^2 \, d\mu \\ &= \int_L (2-\rho^2)\rho^2 d\mu + \int_L 2H^2\rho^2 + (4H^2 - S)H^2 - 2H^2 + S - \tfrac{1}{2} S^2 \, d\mu \\ &= \int_L (2-\rho^2)\rho^2 d\mu + \int_L H^2 S - 2H^2 + S - \tfrac{1}{2} S^2 \, d\mu \\ &= \int_L (2-\rho^2)\rho^2 d\mu + \int_L H^2 (S-2) + \tfrac{1}{2} S(2-S) \, d\mu \\ &= \int_L (2-\rho^2)\rho^2 + (2-S)\left(\tfrac{1}{2} S - H^2\right) d\mu \\ &= \int_L \rho^2(2-\rho^2) + \tfrac{1}{2}\rho^2(2-S) \, d\mu \\ &= \int_L \tfrac{3}{2}\rho^2(2-S) + 2H^2\rho^2 \, d\mu, \end{aligned}$$

where in the second equality we used the Gauss equation $2K = 2 + 4H^2 - S$.

Therefore we obtain the desired integral inequality

$$\int_L \rho^2\bigl(3 - \tfrac{3}{2}S + 2H^2\bigr)\, d\mu \le 0.$$

Particularly if $0 \le S \le 2$, we must have $\rho^2 = 0$ and L is totally umbilic or $\rho^2 \ne 0$, which implies $S = 2$, $H = 0$ and L is a flat minimal Legendrian torus. Thus we have proved Theorem 1.8.

Appendix

In this section we prove the following lemma.

Lemma A.1. *Let L be a Legendrian surface in $\mathbb{S}^5$, and assume that $|\nabla^T h|^2$ and $|\nabla^T H|^2$ are defined as in Lemma 3.8. Then we have*

$$|\nabla^T h|^2 \ge 3|\nabla^T H|^2.$$

Proof. We construct the flowing symmetric tracefree tensor:

$$F^s_{ijk} = h^s_{ijk} - \frac{1}{2}(H^s_i\delta_{jk} + H^s_j\delta_{ik} + H^s_k\delta_{ji}). \tag{A-27}$$

Then it is easy to see that

$$|F|^2 = |\nabla^T h|^2 - 3|\nabla^T H|^2,$$

and we get $|\nabla^T h|^2 \ge 3|\nabla^T H|^2$. □

Final discussions. To end this paper we propose several questions which we will study in the future.

Problem 1. Is any umbilical contact stationary Legendrian surface in $\mathbb{S}^5$ with $0 \le S \le 2$ totally geodesic?

Problem 2. Assume that L is a closed csL submanifold in $\mathbb{S}^{2n+1}$, satisfying $0 \le S \le n$, then is L totally geodesic or $S = n$?

Problem 3. Is any contact stationary Legendrian surface in $\mathbb{S}^5$ with second fundamental form of constant length minimal?

Problem 4. What is the second gap for minimal Legendrian submanifolds in a sphere?

Acknowledgement

I would like to thank professor Guofang Wang for a lot of discussions on Sasakian geometry and useful suggestions during the preparation of this paper and thank professor Ildefonso Castro for his comments and interests in this paper. Many thanks to professor Toru Sasahara for pointing out an error in Lemma 3.3. The author is partially supported by the NSF of China (No.11501421).

References

[Benko et al. 1979] K. Benko, M. Kothe, K.-D. Semmler, and U. Simon, "Eigenvalues of the Laplacian and curvature", *Colloq. Math.* **42**:1 (1979), 19–31. MR Zbl

[Blair 2002] D. E. Blair, *Riemannian geometry of contact and symplectic manifolds*, Progress in Mathematics **203**, Birkhäuser, Boston, 2002. MR Zbl

[Butscher 2009] A. Butscher, "Equivariant gluing constructions of contact stationary Legendrian submanifolds in $\mathbb{S}^{2n+1}$", *Calc. Var. Partial Differential Equations* **35**:1 (2009), 57–102. MR Zbl

[Butscher and Corvino 2012] A. Butscher and J. Corvino, "Hamiltonian stationary tori in Kähler manifolds", *Calc. Var. Partial Differential Equations* **45**:1-2 (2012), 63–100. MR Zbl

[Castro et al. 2006] I. Castro, H. Li, and F. Urbano, "Hamiltonian-minimal Lagrangian submanifolds in complex space forms", *Pacific J. Math.* **227**:1 (2006), 43–63. MR Zbl

[Chen et al. 2012] Q. Chen, S. Hu, and X. Xu, "Construction of Lagrangian submanifolds in $\mathbb{CP}^n$", *Pacific J. Math.* **258**:1 (2012), 31–49. MR Zbl

[Chern et al. 1970] S. S. Chern, M. do Carmo, and S. Kobayashi, "Minimal submanifolds of a sphere with second fundamental form of constant length", pp. 59–75 in *Functional analysis and related fields* (Chicago, 1968), edited by F. E. Browder, Springer, 1970. MR Zbl

[Dillen and Vrancken 1990] F. Dillen and L. Vrancken, "C-totally real submanifolds of Sasakian space forms", *J. Math. Pures Appl.* (9) **69**:1 (1990), 85–93. MR Zbl

[Dong 2007] Y. Dong, "Hamiltonian-minimal Lagrangian submanifolds in Kaehler manifolds with symmetries", *Nonlinear Anal.* **67**:3 (2007), 865–882. MR Zbl

[Dong and Han 2007] Y. Dong and Y. Han, "Some explicit examples of Hamiltonian minimal Lagrangian submanifolds in complex space forms", *Nonlinear Anal.* **66**:5 (2007), 1091–1099. MR Zbl

[Haskins 2004] M. Haskins, "Special Lagrangian cones", *Amer. J. Math.* **126**:4 (2004), 845–871. MR Zbl

[Hélein and Romon 2002] F. Hélein and P. Romon, "Hamiltonian stationary Lagrangian surfaces in Hermitian symmetric spaces", pp. 161–178 in *Differential geometry and integrable systems* (Tokyo, 2000), edited by M. Guest et al., Contemp. Math. **308**, American Mathematical Society, Providence, RI, 2002. MR Zbl

[Hélein and Romon 2005] F. Hélein and P. Romon, "Hamiltonian stationary tori in the complex projective plane", *Proc. London Math. Soc.* (3) **90**:2 (2005), 472–496. MR Zbl

[Hunter and McIntosh 2011] R. Hunter and I. McIntosh, "The classification of Hamiltonian stationary Lagrangian tori in $\mathbb{CP}^2$ by their spectral data", *Manuscripta Math.* **135**:3-4 (2011), 437–468. MR Zbl

[Iriyeh 2005] H. Iriyeh, "Hamiltonian minimal Lagrangian cones in $\mathbb{C}^m$", *Tokyo J. Math.* **28**:1 (2005), 91–107. MR Zbl

[Joyce et al. 2011] D. Joyce, Y.-I. Lee, and R. Schoen, "On the existence of Hamiltonian stationary Lagrangian submanifolds in symplectic manifolds", *Amer. J. Math.* **133**:4 (2011), 1067–1092. MR Zbl

[Kajigaya 2013] T. Kajigaya, "Second variation formula and the stability of Legendrian minimal submanifolds in Sasakian manifolds", *Tohoku Math. J.* (2) **65**:4 (2013), 523–543. MR Zbl

[Kozlowski and Simon 1984] M. Kozlowski and U. Simon, "Minimal immersions of 2-manifolds into spheres", *Math. Z.* **186**:3 (1984), 377–382. MR Zbl

[Lawson 1969] H. B. Lawson, Jr., "Local rigidity theorems for minimal hypersurfaces", *Ann. of Math.* (2) **89**:1 (1969), 187–197. MR Zbl

[Lê 2004] H.-V. Lê, "A minimizing deformation of Legendrian submanifolds in the standard sphere", *Differential Geom. Appl.* **21**:3 (2004), 297–316. MR Zbl

[Lee 2012] Y.-I. Lee, "The existence of Hamiltonian stationary Lagrangian tori in Kähler manifolds of any dimension", *Calc. Var. Partial Differential Equations* **45**:1-2 (2012), 231–251. MR Zbl

[Li 2001] H. Li, "Willmore hypersurfaces in a sphere", *Asian J. Math.* **5**:2 (2001), 365–377. MR Zbl

[Li 2002a] H. Li, "Willmore submanifolds in a sphere", *Math. Res. Lett.* **9**:5-6 (2002), 771–790. MR Zbl

[Li 2002b] H. Li, "Willmore surfaces in $\mathbb{S}^n$", *Ann. Global Anal. Geom.* **21**:2 (2002), 203–213. MR Zbl

[Ludden et al. 1975] G. D. Ludden, M. Okumura, and K. Yano, "A totally real surface in $\mathbb{CP}^2$ that is not totally geodesic", *Proc. Amer. Math. Soc.* **53**:1 (1975), 186–190. MR Zbl

[Ma 2005] H. Ma, "Hamiltonian stationary Lagrangian surfaces in $\mathbb{CP}^2$", *Ann. Global Anal. Geom.* **27**:1 (2005), 1–16. MR Zbl

[Ma and Schmies 2006] H. Ma and M. Schmies, "Examples of Hamiltonian stationary Lagrangian tori in $\mathbb{CP}^2$", *Geom. Dedicata* **118** (2006), 173–183. MR Zbl

[Mironov 2003] A. E. Mironov, "On the Hamiltonian-minimal Lagrangian tori in $\mathbb{CP}^2$", *Sibirsk. Mat. Zh.* **44**:6 (2003), 1324–1328. In Russian; translated in *Siberian Math. J.* **44**:6 (2003), 1039–1042. MR Zbl

[Mironov 2004] A. E. Mironov, "New examples of Hamilton-minimal and minimal Lagrangian submanifolds in $\mathbb{C}^n$ and $\mathbb{CP}^n$", *Mat. Sb.* **195**:1 (2004), 89–102. In Russian; translated in *Sb. Math.* **195**:1-2 (2004), 85–96. MR Zbl

[Mironov 2008] A. E. Mironov, "Spectral data for Hamiltonian-minimal Lagrangian tori in $\mathbb{CP}^2$", *Tr. Mat. Inst. Steklova* **263**:Geometriya, Topologiya i Matematicheskaya Fizika. I (2008), 120–134. In Russian; translated in *Proc. Steklov Inst. Math.* **263**:1 (2008), 112–126. MR Zbl

[Oh 1990] Y.-G. Oh, "Second variation and stabilities of minimal Lagrangian submanifolds in Kähler manifolds", *Invent. Math.* **101**:2 (1990), 501–519. MR Zbl

[Oh 1993] Y.-G. Oh, "Volume minimization of Lagrangian submanifolds under Hamiltonian deformations", *Math. Z.* **212**:2 (1993), 175–192. MR Zbl

[Ono 2005] H. Ono, "Second variation and Legendrian stabilities of minimal Legendrian submanifolds in Sasakian manifolds", *Differential Geom. Appl.* **22**:3 (2005), 327–340. MR Zbl

[Reckziegel 1988] H. Reckziegel, "A correspondence between horizontal submanifolds of Sasakian manifolds and totally real submanifolds of Kählerian manifolds", pp. 1063–1081 in *Topics in differential geometry, II* (Debrecen, 1984), edited by J. Szenthe and L. Tamássy, Colloq. Math. Soc. János Bolyai **46**, North-Holland, Amsterdam, 1988. MR Zbl

[Schoen and Wolfson 2001] R. Schoen and J. Wolfson, "Minimizing area among Lagrangian surfaces: the mapping problem", *J. Differential Geom.* **58**:1 (2001), 1–86. MR Zbl

[Simons 1968] J. Simons, "Minimal varieties in Riemannian manifolds", *Ann. of Math.* (2) **88** (1968), 62–105. MR Zbl

[Smoczyk 2003] K. Smoczyk, "Closed Legendre geodesics in Sasaki manifolds", *New York J. Math.* **9** (2003), 23–47. MR Zbl

[Wu 1988] H. H. Wu, "The Bochner technique in differential geometry", *Math. Rep.* **3**:2 (1988), i–xii and 289–538. MR Zbl

[Yamaguchi et al. 1976] S. Yamaguchi, M. Kon, and Y. Miyahara, "A theorem on C-totally real minimal surface", *Proc. Amer. Math. Soc.* **54** (1976), 276–280. MR Zbl

[Yau 1974] S. T. Yau, "Submanifolds with constant mean curvature, I", *Amer. J. Math.* **96**:2 (1974), 346–366. MR Zbl

Received November 17, 2016.

YONG LUO
SCHOOL OF MATHEMATICS AND STATISTICS
WUHAN UNIVERSITY
WUHAN
CHINA
yongluo@whu.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.121

IRREDUCIBILITY OF THE MODULI SPACE OF STABLE VECTOR BUNDLES OF RANK TWO AND ODD DEGREE ON A VERY GENERAL QUINTIC SURFACE

NICOLE MESTRANO AND CARLOS SIMPSON

The moduli space $M(c_2)$ of stable rank-two vector bundles of degree one on a very general quintic surface $X \subset \mathbb{P}^3$ is irreducible for all $c_2 \geq 4$ and empty otherwise. On the other hand, for a very general sextic surface, the moduli space at $c_2 = 11$ has at least two irreducible components.

1. Introduction

Let $X \subset \mathbb{P}^3_{\mathbb{C}}$ be a very general quintic hypersurface. Let $M(c_2) := M_X(2, 1, c_2)$ denote the moduli space [Huybrechts and Lehn 1997] of stable rank 2 vector bundles on X of degree 1 with $c_2(E) = c_2$. Let $\overline{M}(c_2) := \overline{M}_X(2, 1, c_2)$ denote the moduli space of stable rank 2 torsion-free sheaves on X of degree 1 with $c_2(E) = c_2$. Recall that $\overline{M}(c_2)$ is projective, and $M(c_2) \subset \overline{M}(c_2)$ is an open set, whose complement is called the *boundary*. Let $\overline{M(c_2)}$ denote the closure of $M(c_2)$ inside $\overline{M}(c_2)$. This might be a strict inclusion, as will in fact be the case for $c_2 \leq 10$.

In [Mestrano and Simpson 2011] we showed that $M(c_2)$ is irreducible for $4 \leq c_2 \leq 9$, and empty for $c_2 \leq 3$. In [Mestrano and Simpson 2013] we showed that the open subset $M(10)^{\mathrm{sn}} \subset M(10)$ of bundles with seminatural cohomology is irreducible. Nijsse [1995] showed that $M(c_2)$ is irreducible for $c_2 \geq 16$.

In the present paper, we complete the proof of irreducibility for the remaining intermediate values of c_2.

Theorem 1.1. *For any $c_2 \geq 4$, the moduli space of bundles $M(c_2)$ is irreducible.*

For $c_2 \geq 11$, the moduli space of torsion-free sheaves $\overline{M}(c_2)$ is irreducible. On the other hand, $\overline{M}(10)$ has two irreducible components: the closure $\overline{M(10)}$ of the irreducible open set $M(10)$; and the smallest stratum $M(10, 4)$ of the double dual

This research project was initiated on our visit to Japan, supported by JSPS Grant-in-Aid for Scientific Research (S-19104002). This research was supported in part by French ANR grants G-FIB (BLAN-08-3-352054), SEDIGA (BLAN-08-1-309225), HODAG (09-BLAN-0151-02) and TOFIGROU (933R03/13ANR002SRAR). We thank the University of Miami for hospitality during the completion of this work.

MSC2010: primary 14D20; secondary 14H50, 14J29.

Keywords: vector bundle, surface, moduli space, deformation, boundary.

stratification corresponding to torsion-free sheaves whose double dual has $c_2' = 4$. *Similarly* $\overline{M}(c_2)$ *has several irreducible components when* $5 \le c_2 \le 9$, *too.*

The moduli space $\overline{M}(c_2)$ *is good for* $c_2 \ge 10$, *generically smooth of the expected dimension* $4c_2 - 20$, *whereas for* $4 \le c_2 \le 9$, *the moduli space* $M(c_2)$ *is not good. For* $c_2 \le 3$ *it is empty.*

Yoshioka [1997; 1999; 2001], Gómez [1997] and others have shown that the moduli space of stable torsion-free sheaves with irreducible Mukai vector (which contains, in particular, the case of bundles of rank 2 and degree 1) is irreducible, over an abelian or K3 surface. Those results use the triviality of the canonical bundle, leading to a symplectic structure and implying among other things that the moduli spaces are smooth [Mukai 1984]. Notice that the case of K3 surfaces includes degree 4 hypersurfaces in $\mathbb{P}^3$.

We were motivated to look at a next case, of bundles on a quintic or degree 5 hypersurface in $\mathbb{P}^3$ where $K_X = \mathcal{O}_X(1)$ is ample but not by very much. This paper is the third in a series starting with [Mestrano and Simpson 2011; 2013] dedicated to Professor Maruyama who, along with Gieseker, pioneered the study of moduli of bundles on higher dimensional varieties [Gieseker 1977; 1988; Maruyama 1973; 1975; 1982]. Recall that the moduli space of stable bundles is irreducible for $c_2 \gg 0$ on any smooth projective surface [Gieseker and Li 1994; Li 1993; O'Grady 1993; 1996], but there exist surfaces, such as smooth hypersurfaces in $\mathbb{P}^3$ of sufficiently high degree [Mestrano 1997], where the moduli space is not irreducible for intermediate values of c_2.

Our theorem shows that the irreducibility of the moduli space of bundles $M(c_2)$, for all values of c_2, can persist into the range where K_X is ample. On the other hand, the fact that $\overline{M}(10)$ has two irreducible components, means that if we consider all torsion-free sheaves, then the property of irreducibility in the good range has already started to fail in the case of a quintic hypersurface.

We furthermore show in Section 11 below that irreducibility fails for stable vector bundles on surfaces of degree $d = 6$. This improves the result of [Mestrano 1997] where nonirreducibility had been obtained on surfaces of degree $d \ge 27$.

A possible application of the irreducibility theorem to the case of Calabi–Yau varieties could be envisioned by noting that a general hyperplane section of a quintic threefold in $\mathbb{P}^4$ will be a quintic surface $X \subset \mathbb{P}^3$.

Outline of the proof. The starting point is O'Grady's [1993; 1996] method of deformation to the boundary, as exploited by Nijsse [1995] in the case of a very general quintic hypersurface. We use in particular some of the intermediate results of Nijsse who showed, for example, that $\overline{M}(c_2)$ is connected for $c_2 \ge 10$.

Application of these techniques is made possible by the explicit description of the moduli spaces $M(c_2)$ for $4 \leq c_2 \leq 9$ and the partial result for $M(10)$ obtained in [Mestrano and Simpson 2011; 2013].

Our approach therefore has a botanical flavor. The information gleaned from the descriptions in [Mestrano and Simpson 2011] allows us to understand the boundary components. It turns out that the bigger components growing out of these will correspond to bundles with seminatural cohomology, so that the result of [Mestrano and Simpson 2013] applies. We should stress that it is not a priori clear something like this should happen — the possibility of getting to the proof in this way becomes accessible only through an understanding of the components at lower levels. This will present a challenge for generalization to other surfaces.

The boundary $\partial \overline{M}(c_2) := \overline{M}(c_2) - M(c_2)$ is the set of points corresponding to torsion-free sheaves which are not locally free. We just endow $\partial \overline{M}(c_2)$ with its reduced scheme structure. There might in some cases be a better nonreduced structure which one could put on the boundary or onto some strata, but that won't be necessary for our argument and we don't worry about it here.

We can further refine the decomposition

$$\overline{M}(c_2) = M(c_2) \sqcup \partial \overline{M}(c_2)$$

by the *double dual stratification* [O'Grady 1996]. Let $M(c_2; c_2')$ denote the locally closed subset, again with its reduced scheme structure, parametrizing sheaves F which fit into an exact sequence

$$0 \to F \to F^{**} \to S \to 0$$

such that $F \in \overline{M}(c_2)$ and S is a coherent sheaf of finite length $d = c_2 - c_2'$ hence $c_2(F^{**}) = c_2'$. Notice that $E = F^{**}$ is also stable so it is a point in $M(c_2')$. The stratum can be nonempty only when $c_2' \geq 4$, which shows by the way that $\overline{M}(c_2)$ is empty for $c_2 \leq 3$. The boundary now decomposes into locally closed subsets

$$\partial \overline{M}(c_2) = \coprod_{4 \leq c_2' < c_2} M(c_2; c_2').$$

Let $\overline{M(c_2, c_2')}$ denote the closure of $M(c_2, c_2')$ in $\overline{M}(c_2)$. Notice that we don't know anything about the position of this closure with respect to the stratification; its boundary will not in general be a union of strata. We can similarly denote by $\overline{M(c_2)}$ the closure of $M(c_2)$ inside $\overline{M}(c_2)$, a subset which might well be strictly smaller than $\overline{M}(c_2)$.

The construction $F \mapsto F^{**}$ provides, by the definition of the stratification, a well-defined map

$$M(c_2; c_2') \to M(c_2').$$

The fiber over $E \in M(c_2')$ is the Grothendieck Quot-scheme $\mathrm{Quot}(E; d)$ of quotients of E of length $d := c_2 - c_2'$.

It follows from Li's theorem [Li 1993, Proposition 6.4] that if $M(c_2')$ is irreducible, then $M(c_2; c_2')$ and hence $\overline{M(c_2; c_2')}$ are irreducible, with $\dim(M(c_2; c_2')) = \dim(M(c_2')) + 3(c_2 - c_2')$. See Corollary 4.3 below. From the previous papers [Mestrano and Simpson 2011; 2013], we know the dimensions of $M(c_2')$, so we can fill in the dimensions of the strata, as will be summarized in Table 2. Furthermore, by [Mestrano and Simpson 2011] and Li's theorem, the strata $M(c_2; c_2')$ are irreducible whenever $c_2' \le 9$.

Nijsse [1995] proves that $\overline{M}(c_2)$ is connected whenever $c_2 \ge 10$, using O'Grady's [1993; 1996] techniques. This is discussed and we review the proof in [Mestrano and Simpson 2016]. By [Mestrano and Simpson 2011], the moduli space $\overline{M}(c_2)$ is *good*, that is to say it is generically reduced of the expected dimension $4c_2 - 20$, whenever $c_2 \ge 10$. In particular, the dimension of the Zariski tangent space, minus the dimension of the space of obstructions, is equal to the dimension of the moduli space. The Kuranishi theory of deformation spaces implies that $\overline{M}(c_2)$ is locally a complete intersection. Hartshorne's [1962] connectedness theorem now says that if two different irreducible components of $\overline{M}(c_2)$ meet at some point, then they intersect in a codimension 1 subvariety. This intersection has to be contained in the singular locus.

The singular locus in $M(c_2)$ contains a subvariety denoted $V(c_2)$, which is the set of bundles E with $h^0(E) > 0$. It is the image of the space Σ_{c_2} of extensions

$$0 \to \mathcal{O}_X \to E \to J_P(1) \to 0,$$

where P satisfies Cayley–Bacharach for quadrics. For $c_2 \ge 10$, $V(c_2)$ is irreducible of dimension $3c_2 - 11$. For $c_2 \ge 11$ one can see directly that the closure of $V(c_2)$ meets the boundary. For $c_2 = 10$, bundles in $V(10)$ almost have seminatural cohomology, in the sense that any deformation moving away from $V(10)$ will have seminatural cohomology, so $V(10)$ is contained only in the irreducible component constructed in [Mestrano and Simpson 2013], and that component meets the boundary. On the other hand, any other irreducible components of the singular locus have strictly smaller dimension [Mestrano and Simpson 2011, Corollary 7.1].

These properties of the singular locus, together with the connectedness statement of [Nijsse 1995], allow us to show that any irreducible component of $\overline{M}(c_2)$ meets the boundary. O'Grady proves furthermore an important lemma, that the intersection with the boundary must have pure codimension 1.

We explain the strategy for proving irreducibility of $M(10)$ and $M(11)$ below, but it will perhaps be easiest to explain first why this implies irreducibility of $M(c_2)$ for $c_2 \ge 12$. Based on O'Grady's method, this is the same strategy as was used by Nijsse who treated the cases $c_2 \ge 16$.

Suppose $c_2 \geq 12$ and $Z \subset \overline{M}(c_2)$ is an irreducible component. Suppose inductively we know that $M(c_2 - 1)$ is irreducible. Then $\partial Z := Z \cap \partial\overline{M}(c_2)$ is a nonempty subset in Z of codimension 1, thus of dimension $4c_2 - 21$. However, by looking at Table 2, the boundary $\partial\overline{M}(c_2)$ is a union of the stratum $M(c_2, c_2 - 1)$ of dimension $4c_2 - 21$, plus other strata of strictly smaller dimension. Therefore, ∂Z must contain $M(c_2, c_2 - 1)$. But, the general torsion-free sheaf parametrized by a point of $M(c_2, c_2 - 1)$ is the kernel F of a general surjection $E \to S$ from a stable bundle E general in $M(c_2 - 1)$, to a sheaf S of length 1. We claim that F is a smooth point of the moduli space $\overline{M}(c_2)$. Indeed, if F were a singular point then there would exist a nontrivial coobstruction $\phi : F \to F(1)$; see [Langer 2008; Mestrano and Simpson 2011; Zuo 1991]. This would have to come from a nontrivial coobstruction $E \to E(1)$ for E, but that cannot exist because a general E is a smooth point since $M(c_2 - 1)$ is good. Thus, F is a smooth point of the moduli space. It follows that a given irreducible component of $M(c_2, c_2 - 1)$ is contained in at most one irreducible component of $\overline{M}(c_2)$. On the other hand, by the induction hypothesis $M(c_2 - 1)$ is irreducible, so $M(c_2, c_2 - 1)$ is irreducible. This gives the induction step, that $M(c_2)$ is irreducible.

The strategy for $M(10)$ is similar. However, due to the fact that the moduli spaces $M(c_2')$ are not good for $c_2' \leq 9$, in particular they tend to have dimensions bigger than the expected dimensions, there are several boundary strata which can come into play. Luckily, we know that the $M(c_2')$, hence all of the strata $M(10, c_2')$, are irreducible for $c_2' \leq 9$.

The dimension of $M(10)$, equal to the expected one, is 20. Looking at the row $c_2 = 10$ in Table 2 below, one may see that there are three strata $M(10, 9)$, $M(10, 8)$ and $M(10, 6)$ with dimension 19. These can be irreducible components of the boundary ∂Z if we follow the previous argument. More difficult is the case of the stratum $M(10, 4)$ which has dimension 20. A general point of $M(10, 4)$ is not in the closure of $M(10)$, in other words $M(10, 4)$, which is closed since it is the lowest stratum, constitutes a separate irreducible component of $\overline{M}(10)$. Now, if $Z \subset M(10)$ is an irreducible component, ∂Z could contain a codimension 1 subvariety of $M(10, 4)$.

The next idea is to use the main result of [Mestrano and Simpson 2013], that the moduli space $M(10)^{\text{sn}}$ of bundles with seminatural cohomology, is irreducible. To prove that $M(10)$ is irreducible, it therefore suffices to show that a general point of any irreducible component Z, has seminatural cohomology. From [Mestrano and Simpson 2013] there are two conditions that need to be checked: $h^0(E) = 0$ and $h^1(E(1)) = 0$. The first condition is automatic for a general point, since the locus $V(10)$ of bundles with $h^0(E) > 0$ has dimension $3 \cdot 10 - 11 = 19$ so cannot contain a general point of Z. For the second condition, it suffices to note that a general sheaf F in any of the strata $M(10, 9)$, $M(10, 8)$ and $M(10, 6)$ has $h^1(F(1)) = 0$;

and to show that the subspace of sheaves F in $M(10,4)$ with $h^1(F(1)) > 0$ has codimension ≥ 2. This latter result is treated in Section 7, using the dimension results of Ellingsrud and Lehn for the scheme of quotients of a locally free sheaf, generalizing Li's theorem. This is how we will show irreducibility of $M(10)$.

The full moduli space of torsion-free sheaves $\overline{M}(10)$ has two different irreducible components, the closure $\overline{M(10)}$ and the lowest stratum $M(10,4)$. This distinguishes the case of the quintic surface from the cases of abelian and K3 surfaces, where the full moduli spaces of stable torsion-free sheaves were irreducible [Yoshioka 1999; 2001; Gómez 1997].

For $M(11)$, the argument is almost the same as for $c_2 \geq 12$. However, there are now two different strata of codimension 1 in the boundary: $M(11,10)$ coming from the irreducible variety $M(10)$, and $M(11,4)$ which comes from the other 20-dimensional component $M(10,4)$ of $\overline{M}(10)$. To show that these two can give rise to at most a single irreducible component in $M(11)$, completing the proof, we will note that they do indeed intersect, and furthermore that the intersection contains smooth points.

After the end of the proof of Theorem 1.1, the last two sections of the paper treat some related considerations.

In Section 10 we provide a correction and improvement to [Mestrano and Simpson 2011, Lemma 5.1] and answer that paper's Question 5.1. Recall from there that a coobstruction may be interpreted as a sort of Higgs field with values in the canonical bundle K_X; it has a spectral surface $Z \subset \mathrm{Tot}(K_X)$. The question was to bound the irregularity of a resolution of singularities of the spectral surface Z. We show in Lemma 10.1 that the irregularity vanishes.

Example on a sextic. At the end of the paper in Section 11, we show Theorem 1.1 is sharp as far as the degree 5 of the very general hypersurface is concerned. In the case of bundles on very general hypersurfaces X^6 of degree 6, we show in Theorem 11.4 that the moduli space $M_{X^6}(2,1,11)$ of stable rank two bundles of degree 1 and $c_2 = 11$ has at least two irreducible components. This improves the result of [Mestrano 1997], bringing from 27 down to 6 the degree of a very general hypersurface on which there exist two irreducible components. We expect that there will be several irreducible components in any degree ≥ 6 but that isn't shown here.

2. Preliminary facts

The moduli space $\overline{M}(c_2)$ is locally a fine moduli space. The obstruction to existence of a Poincaré universal sheaf on $\overline{M}(c_2) \times X$ is an interesting question but not considered in the present paper. A universal family exists étale-locally over $\overline{M}(c_2)$ so for local questions we may consider $\overline{M}(c_2)$ as a fine moduli space.

c_2	4	5	6	7	8	9
$\dim(M)$	2	3	7	9	13	16
$\dim(\mathrm{obs})$	6	3	3	3	1	1
$h^1(E(1))$	0	1	0	0	0	0
generically	sm	sm	sm	nr	sm	nr

Table 1. Moduli spaces for $c_2 \le 9$.

The Zariski tangent space to $\bar{M}(c_2)$ at a point E is $\mathrm{Ext}^1(E, E)$. If E is locally free, this is the same as $H^1(\mathrm{End}(E))$. The *space of obstructions* $\mathrm{obs}(E)$ is by definition the kernel of the surjective map

$$\mathrm{Tr} : \mathrm{Ext}^2(E, E) \to H^2(\mathcal{O}_X).$$

The *space of coobstructions* is the dual $\mathrm{obs}(E)^*$ which is, by Serre duality with $K_X = \mathcal{O}_X(1)$, equal to $\mathrm{Hom}^0(E, E(1))$, the space of maps $\phi : E \to E(1)$ such that $\mathrm{Tr}(\phi) = 0$ in $H^0(\mathcal{O}_X(1)) \cong \mathbb{C}^4$. Such a map is called a *coobstruction*.

Since a torsion-free sheaf E of rank two and odd degree can have no rank-one subsheaves of the same slope, all semistable sheaves are stable, and Gieseker and slope stability are equivalent. If E is a stable sheaf then $\mathrm{Hom}(E, E) = \mathbb{C}$ so the space of trace-free endomorphisms is zero. Notice that $H^1(\mathcal{O}_X) = 0$ so we may disregard the trace-free condition for $\mathrm{Ext}^1(E, E)$. An Euler characteristic calculation gives

$$\dim(\mathrm{Ext}^1(E, E)) - \dim(\mathrm{obs}(E)) = 4c_2 - 20,$$

and this is called the *expected dimension* of the moduli space. The moduli space is said to be *good* if the dimension is equal to the expected dimension.

Lemma 2.1. *If the moduli space is good, then it is locally a complete intersection.*

Proof. Kuranishi theory expresses the local analytic germ of the moduli space $\bar{M}(c_2)$ at E, as $\Phi^{-1}(0)$ for a holomorphic map of germs $\Phi : (\mathbb{C}^a, 0) \to (\mathbb{C}^b, 0)$ where $a = \dim(\mathrm{Ext}^1(E, E))$ (resp. $b = \dim(\mathrm{obs}(E))$). Hence, if the moduli space has dimension $a - b$, it is a locally complete intersection. □

We investigated closely the structure of the moduli space for $c_2 \le 9$ in [Mestrano and Simpson 2011].

Proposition 2.2. *The moduli space $M(c_2)$ is empty for $c_2 \le 3$. For $4 \le c_2 \le 9$, the moduli space $M(c_2)$ is irreducible. It has dimension strictly bigger than the expected one, for $4 \le c_2 \le 8$, and for $c_2 = 9$ it is generically nonreduced but with dimension equal to the expected one; it is also generically nonreduced for $c_2 = 7$. The dimensions of the moduli spaces, the dimensions of the spaces of obstructions at a general point, and the dimensions $h^1(E(1))$ for a general bundle E in $M(c_2)$, are given in Table 1 above.*

The proof of Proposition 2.2 will be given in the next section, with a review of the cases $c_2 \leq 9$ from the paper [Mestrano and Simpson 2011].

We also proved that the moduli space is good for $c_2 \geq 10$, known by Nijsse [1995] for $c_2 \geq 13$.

Proposition 2.3. *For $c_2 \geq 10$, the moduli space $M(c_2)$ is good. The singular locus $M(c_2)^{\mathrm{sing}}$ is the union of the locus $V(c_2)$ consisting of bundles with $h^0(E) > 0$, which has dimension $3c_2 - 11$, plus other pieces of dimension ≤ 13 which in particular have codimension ≥ 6.*

Proof. Following O'Grady's and Nijsse's terminology $V(c_2)$ denotes the locus which is the image of the moduli space of bundles together with a section, called Σ_{c_2} or sometimes $\{E, P\}$. See [Mestrano and Simpson 2011, Theorem 7.1]. Any pieces of the singular locus corresponding to bundles which are not in $V(c_2)$, have dimension ≤ 13 by [Mestrano and Simpson 2011, Corollary 5.1] (see Lemma 10.1 below for a correction and improvement of this statement). □

The case $c_2 = 10$ is an important central point in the classification, where the case-by-case treatment gives way to a general picture. In [Mestrano and Simpson 2013], we proved the following partial result that will be used in the present paper to complete the proof of irreducibility.

Proposition 2.4. *Let $M(10)^{\mathrm{sn}} \subset M(10)$ denote the open subset of bundles $E \in M(10)$ which have* ***seminatural cohomology****, that is, where for any m at most one of $h^i(E(m))$ is nonzero for $i = 0, 1, 2$. Then $E \in M(10)^{\mathrm{sn}}$ if and only if $h^0(E) = 0$ and $h^1(E(1)) = 0$. The moduli space $M(10)^{\mathrm{sn}}$ is irreducible.*

Proof. See [Mestrano and Simpson 2013], Theorem 0.2 and Corollary 3.5. □

3. Review of $c_2 \leq 9$

Our strategy of proof uses in a fundamental way an understanding of the irreducible components for $c_2 \leq 9$ that were studied in [Mestrano and Simpson 2011]. The discussion of these moduli spaces went by a sometimes exhaustive classification of cases Lemmas 7.3, 7.4 there. In retrospect we can give more uniform proofs of some parts. For this reason, and for the reader's convenience, it is worthwhile to review here some of the arguments leading to the proof of Proposition 2.2. This section may, however, be skipped or perused lightly on the first reading.

There is a change of notation with respect to that work. There we considered bundles of degree -1. The bundle of degree 1 denoted here by E is the same as the bundle denoted by $E(1)$ there. Thus Lemma 5.2 there speaks of $h^1(E)$ in our notation. The present notation was already in effect in [Mestrano and Simpson 2013]. Fortunately, the indexing by second Chern class remains the same in both cases.

Following O'Grady, we denote by $V(c_2) \subset M(c_2)$ the subvariety of bundles such that $h^0(E) > 0$. For $c_2 \le 9$ the Euler characteristic argument of [Mestrano and Simpson 2011, §6.1] tells us that $h^0(E) > 0$ for any E, so $V(c_2)$ is the full moduli space.

It will be useful to consider the moduli space Σ_{c_2} consisting of pairs (E, η) where $E \in M(c_2)$ and $\eta \in H^0(E)$ is a nonzero section. The pairs are taken up to isomorphism, i.e., up to scaling of the section, so the fiber of the map $\Sigma_{c_2} \to V(c_2)$ over a bundle E is the projective space $\mathbb{P}H^0(E)$.

Each irreducible component of Σ_{c_2} has dimension $\ge 3c_2 - 11$, see [O'Grady 1996; Nijsse 1995] or [Mestrano and Simpson 2011, Corollary 3.1].

A point of Σ_{c_2} may also be considered as an extension of the form

$$0 \to \mathcal{O}_X \to E \to J_{P/X}(1) \to 0,$$

again up to isomorphism. We therefore employ the notation $\{E, P\} := \Sigma_{c_2}$, too.

Such an extension exists, with E a bundle, if and only if $P \subset X$ is locally a complete intersection of length c_2 and satisfies the Cayley–Bacharach condition for quadrics denoted CB(2). See [Barth 1977; Griffiths and Harris 1978; Reider 1988] and the references for the Hartshorne–Serre correspondence discussed in [Arrondo 2007] for the origins of this principle.

Denote by $\{P\}$ the Hilbert scheme of l.c.i. subschemes P that satisfy CB(2). The map $\{E, P\} \to \{P\}$ has fibers described as follows: the fiber over P is a dense open subset[1] of the projective space of all extensions $\mathbb{P}\,\mathrm{Ext}^1(J_{P/X}(1), \mathcal{O}_X)$; its dimension by duality is $h^1(J_{P/X}(1)) - 1$.

Consider c the number of conditions imposed by P on quadrics. This is related to $h^1(E(1))$ by the exact sequences

$$H^0(\mathcal{O}_X(2)) \to H^0(\mathcal{O}_P(2)) \to H^1(J_{P/X}(2)) \to 0$$

and

$$0 \to H^1(E(1)) \to H^1(J_{P/X}(2)) \to H^2(\mathcal{O}_X(1)) \to 0,$$

where $H^2(E(1)) = H^0(E(1))^* = 0$ by stability, and $H^2(\mathcal{O}_X(1)) = H^2(K_X) = \mathbb{C}$. The number c is the rank of the evaluation map of $H^0(\mathcal{O}_X(2))$ on P, so $h^1(J_{P/X}(2)) = c_2 - c$, and by the second exact sequence we have $h^1(E(1)) = c_2 - c - 1$.

The number $c_2 - c - 1$ is also equal to the dimension of the fiber of the map from the space of extensions $\{E, P\}$ to the Hilbert scheme of subschemes $\{P\}$. As stated previously, the space of extensions $\{E, P\}$ fibers over the moduli space of bundles $\{E\}$ with fiber $\mathbb{P}H^0(E)$ of dimension $h^1(J_{P/X}(1))$.

[1] It is the open subset of extensions such that E is locally free, nonempty because of the conditions on P.

The locus $V(c_2)$, image of Σ_{c_2}, is the main piece of the set of potentially obstructed bundles, that is to say bundles for which the space of obstructions is nonzero.

The other pieces are of smaller dimension. There was an error in the proof of this dimension estimate, Lemma 5.1 and hence Corollary 5.1 in [Mestrano and Simpson 2011]. These will be corrected and improved in a separate section at the end of the present paper, see Lemma 10.1 below.

Using the Cayley–Bacharach condition. Recall that a 0-dimensional subscheme $P \subset \mathbb{P}^3$ satisfies the *Cayley–Bacharach condition* $\mathrm{CB}(n)$ if, for any subscheme $P' \subset P$ with length $\ell(P') = \ell(P) - 1$, a section $f \in H^0(\mathcal{O}_{\mathbb{P}^3}(n))$ vanishing on P' must also vanish on P. When $P \subset X$ this is the condition governing the existence of an extension of $J_{P/X}(n-1)$ by $\mathcal{O}_X$ that is locally free. For the study of Σ_{c_2} we are therefore interested in subschemes satisfying $\mathrm{CB}(2)$.

See [Mestrano and Simpson 2011; 2013] and the survey [Mestrano and Simpson 2016] for details on the basic techniques we use to analyze the Cayley–Bacharach condition.

If $U \subset \mathbb{P}^3$ is a divisor, usually for us a plane, and P a subscheme, there is a *residual subscheme* P' for P with respect to U. In the case of distinct points it is just the complement of $P \cap U$, but more generally it has a schematic meaning with $\ell(P') + \ell(P \cap U) = \ell(P)$. If P satisfies $\mathrm{CB}(n)$ and U has degree m then the residual P' satisfies $\mathrm{CB}(n-m)$.

The following fact will be used often: if P' is the residual of P with respect to U, and if $Z \subset \mathbb{P}^3$ is a subvariety, then the length of $Z \cap P$ at any point is at least equal to the length of $Z \cap P'$. So for example if P' has 3 points in a line (schematically), then P does too.

It is easy to see that the Cayley–Bacharach condition $\mathrm{CB}(2)$ cannot be satisfied by ≤ 3 points, so the moduli space is empty for $c_2 \leq 3$. Here is a case-by-case review of the cases $4 \leq c_2 \leq 9$.

For $c_2 = 4, 5$. Here the subscheme P is either 4 or 5 points contained in a line. Both of these configurations impose $c = 3$ conditions on quadrics, since $h^0(\mathcal{O}_{\mathbb{P}^1}(2)) = 3$. This gives values of $4 - 3 - 1 = 0$ and $5 - 3 - 1 = 1$ for $h^1(E(1))$ respectively. The moduli space is generically smooth and its dimension is equal to $c_2 - 2$ by [Mestrano and Simpson 2011, Lemma 7.7]. This may be seen directly from the more explicit descriptions we shall give in Section 7 below. We get the dimension of the space of coobstructions by subtracting the expected dimension. This completes the proof of Proposition 2.2 for the columns $c_2 = 4, 5$.

For $c_2 = 6, 7$. In both cases, the Euler characteristic argument of [Mestrano and Simpson 2011, Section 6.1] gives $h^0(E) = 2$, hence $h^0(J_{P/X}(1)) = 1$ and P is contained in a unique plane U. By Lemma 5.5 there, the space of obstructions has dimension 3.

For $c_2 = \ell(P) = 6$, see Proposition 7.4 there that we now review. The number c of conditions imposed on quadrics has to be ≤ 5, in particular P is contained in a planar conic $Y \subset U$. However, $c \leq 4$ may be ruled out by the size of P and the Cayley–Bacharach condition; see the second paragraph of §7.5 there. It follows that the dimension of $\{E, P\}$ equals the dimension of $\{P\}$, and as noted above this dimension is $\geq 3c_2 - 11 = 7$.

Look at the family of length 6 subschemes $P \subset X \cap Y$ such that all points of P are located either at smooth points of Y, or at smooth points of $X \cap U$. Such a subscheme is uniquely determined by its multiplicities at each point, so given Y the set of choices of P is discrete and if we generalize Y, the subscheme P generalizes. Therefore, this defines a set of irreducible components of dimension equal to the dimension of the space of choices of Y, that is 8. For U fixed and Y general, the choice of P is equivalent to the choice of complementary set of 4 points in $Y \cap X$; but since any 4 points in the plane lie on a conic, the monodromy action as we move Y can take any choice of 4 points to any other one. Therefore, this family is a single irreducible component of dimension 8.

The remaining locus of P containing a point where Y is singular and U is tangent to X, has dimension ≤ 5. For example if there is one such point, then the space of choices of U has dimension 2; the space of choices of Y has dimension 2; and by the precise estimate of [Briançon et al. 1981, Proposition 4.3], noting that Y has multiplicity 2 at the singular point, the space of choices of P has dimension ≤ 1. For more points, we get one further dimension of the space of choices of P for each other point but more than 1 new condition imposed by the tangencies. Therefore, the locus of subschemes not fitting into the situation of the previous paragraph has dimension < 7, and it cannot produce a new irreducible component.

This completes the discussion for $c_2 = 6$: we have an irreducible component of $\{E, P\}$ of dimension 8 whose general point consists of a choice of 6 out of the 10 intersection points in $X \cap Y$ for a plane conic Y. Since $h^0(E) = 2$ the dimension of $\{E\}$ is 7. For the table, notice that $h^1(E(1)) = 6-5-1 = 0$. Comparing dimension, expected dimension $4 \cdot 6 - 20 = 4$ and the dimension 3 of the space of obstructions, we find that the moduli space is generically smooth with vanishing obstruction maps.

Consider now the case $c_2 = 7$. See [Mestrano and Simpson 2011, Proposition 7.3] to be reviewed as follows. As previously from the second paragraph of §7.5 the same work, the case $c \leq 4$ may be ruled out. If $c = 5$, then P would be contained in a plane conic $Y \subset U$, but using the same arguments as before the dimension of the space of choices of P would be ≤ 8; however any irreducible component of $\{E, P\}$ has dimension $\geq 3 \cdot 7 - 11 = 10$ and the fiber of the map to $\{P\}$ has dimension 1, so a family of subschemes P of dimension ≤ 8 cannot contribute an irreducible component. Therefore we may suppose $c = 6$, the dimensions of $\{E, P\}$ and $\{P\}$ are the same and are ≥ 10. For a given plane U the space of choices of subscheme $P \subset X \cap U$ of length 7 has dimension 7 by [Briançon et al. 1981]. The space of choices of P such that $U \cap X$ is singular (i.e., U tangent to X), therefore has dimension ≤ 9 and cannot contribute. If U is a plane such that $X \cap U$ is smooth, the Hilbert scheme of $P \subset X \cap U$ is irreducible and a general point corresponds to choosing 7 distinct points. We conclude that $\{E, P\}$ is irreducible of dimension 10 with general point consisting of a general subscheme $P \subset U \cap X$ of length 7 that indeed satisfies CB(2) imposing $c = 6$ conditions on quadrics.

Notice that since $h^0(E) = 2$ the map $\{E, P\} \to \{E\}$ is a fibration with fibers $\mathbb{P}^1$ so the corresponding irreducible component of the moduli space has dimension 9 as filled into the table. At a general point where P imposes $c = 6$ conditions on quadrics, we get $h^1(E(1)) = 7 - 6 - 1 = 0$. From [Mestrano and Simpson 2011, Proposition 7.3], by comparing dimensions the moduli space is generically nonreduced. This treats the column $c_2 = 7$.

For $c_2 = 8$. See the discussion in [Mestrano and Simpson 2011, Section 6.2] and Theorem 7.2 there which will now be reviewed with some improvement in the arguments allowing us to bypass certain case-by-case considerations.

Any component of $\{E, P\}$ has dimension $\geq 3 \cdot 8 - 11 = 13$.

The following technique, involving the residual subscheme recalled above, will be useful.

Lemma 3.1. *Suppose $U \subset \mathbb{P}^3$ is a plane, and let P' denote the residual subscheme for P with respect to U. If nonempty, P' satisfies* CB(1), *so $\ell(P') \geq 3$ and in case of equality P' is collinear.*

Let c be the number of conditions imposed on quadrics in $\mathbb{P}^3$ passing through P, and let n be the number of additional conditions on these quadrics needed to insure their vanishing on U. Suppose $10 - c \geq n + 1$. Then there exists a quadric containing P of the form $U \cup U'$ where U' is another plane, containing P. In particular, $P' \subset U'$. If $10 - c \geq n + 2$ then P' is contained in a line, and if $10 - c \geq n + 3$ then $P \subset U$.

Proof. The first paragraph is a restatement of the basic property of the residual subscheme. Note that one or two points, or three noncollinear points, cannot be CB(1).

In the second paragraph, we could define n as the dimension of the image of

$$H^0(J_{P/\mathbb{P}^3}(2)) \to H^0(\mathcal{O}_U(2)).$$

If $10 - c \geq n + 1$, then it means that we can impose n additional conditions (say, vanishing at general points of U) on the $(10 - c)$-dimensional space quadrics $H^0(J_{P/\mathbb{P}^3}(2))$, to get one that vanishes on U. This quadric has the form $U \cup U'$ of the union of U with another plane U'. By definition the residual is contained in U'. If $10-c \geq n+2$, then the U' move in a 2-dimensional family so they cut out a line containing P'. If $10-c \geq n+3$, the family of U' cuts out a point; however P' satisfying CB(1) cannot be a single point, so in this case it is empty and $P \subset U$. □

Look at the value of c at a general point of an irreducible component. The case $c \leq 5$ may be ruled out (using a simpler version of the subsequent arguments), so we may assume either $c = 6$ or $c = 7$. If $c = 6$ then the fiber of $\{E, P\} \to \{P\}$ has dimension 1 and $\{P\}$ has dimension ≥ 12, whereas if $c = 7$ then the irreducible component of $\{E, P\}$ is the same as that of $\{P\}$, and $\{P\}$ has dimension ≥ 13.

It follows that a general P is not contained in any multiple of a plane. Indeed, the space of the $m.U$ has dimension 3 whereas for any one, the dimension of the space of length 8 subschemes $P \subset X \cap m.U$ is ≤ 8 by [Briançon et al. 1981]. (Here and below, by $m.U$ we denote the m-tuple scheme structure on U.)

Lemma 3.2. *In a given irreducible component, a general P does not contain a collinear subscheme of length ≥ 3 in a line.*

Proof. Start by noting that P is not contained in $U \cup L$ for a plane U and a line L. The space of quadrics containing $U \cup L$ has dimension 2, whereas $c \leq 7$ so there would be a third quadric containing P. One can see that it would have to contain L, so it defines a plane conic $Y \subset U$, meeting L, and $P \subset Y \cup L$. But the dimension of the space of choices of Y, L is 3 for the plane, 5 for the conic, 1 for the intersection point with L and then 2 for the direction of L making 11. Given Y, L the choice of P is discrete (except in some degenerate cases[2]). The set of such P can therefore not be dense in an irreducible component.

We now show that P cannot have three points collinear in a line R, assuming to the contrary that it does. Choose a point $p \in P$ not contained in R (possible by the paragraph above the lemma). Let U be the plane spanned by p and R. Vanishing on $P \cap R$ and at p imposes 4 conditions on conics of U.

In the case $c = 6$, by Lemma 3.1 with $n \leq 2$ so $4 = 10-c \geq n+2$, the residual P' of P with respect to U is contained in a line L, and we get $P \subset U \cup L$, contradicting the first paragraph.

In the case $c = 7$, by Lemma 3.1 with $n \leq 2$, so $3 = 10-c \geq n+1$, we get $P \subset U \subset U'$. Both U and U' must contain points not touching R. The residual P' of P with respect to U has length ≥ 4, indeed if it were to consist of 3 points they would have to be collinear by the CB(1) property but that would give $P \subset U \cup L$.

[2]Since P is not contained in a double plane, Y is not a double line; in the other cases, singularities of $X \cap (Y \cup L)$ are always contained in planar singularities of multiplicity 2 so by [Briançon et al. 1981] the dimension of the space of P increases by 1 at any such point; but existence of the singularity imposes at least one additional condition decreasing the dimension of the space of Y, L.

If U' doesn't contain R, the intersection $P \cap (U' \cup R)$ has length[3] at least 7, but since $U' \cup R$ is cut out by quadrics the CB(2) property of P says that in fact $P \subset (U' \cup R)$ contradicting the first paragraph of the proof.

Suppose $R \subset U'$. Given a residual point lying along R, it cannot correspond to a subscheme leaving R in a direction different from U'. For in that case, we could let U_2 be the plane contacting this direction, different from U or U', and applying Lemma 3.1 again would give $P \subset U_2 \cup U_3$ contradicting the fact that both U and U' contain points of P not on R. So, any point of P' along R corresponds to a point of extra contact with U'. We conclude that the residual subscheme of $P \cap U'$ with respect to $R \subset U'$, has length ≥ 2. Therefore, $n = 1$ conditions suffice to imply vanishing of quadrics on U' so by Lemma 3.1 this time with $3 = 10 - c \geq n + 2$ we find that the residual of P with respect to U' is contained in a line. This again gives P contained in a plane plus a line, contradicting the first paragraph of the proof. □

We may now show that the case $c = 6$ doesn't contribute a general point of an irreducible component. Choose 3 points of P defining a plane U and apply Lemma 3.1 adding $n \leq 3$ extra conditions: we get at least one quadric in our family that has the form $U \cup U'$. Now if $U \cap P$ has length 5, then the residual would have length 3 and satisfy CB(1); therefore it would have to be collinear, contradicting the previous lemma. It follows that $U \cap P$ and $U' \cap P$ both have length 4. But then, it actually sufficed to add $n \leq 2$ conditions so we get a line containing the residual, again contradicting Lemma 3.2. This finishes ruling out the possibility of an irreducible component whose general point imposes $c \leq 6$ conditions on quadrics.

Therefore assume $c = 7$. Now $\{E, P\}$ and $\{P\}$ have the same dimension which is ≥ 13. There is a vector space of dimension $10 - c = 3$ of quadrics passing through P. Let H_1, H_2, H_3 denote the elements of a basis of this space.

Here the proof divides into an analysis of two distinct cases; these were called (a) and (b) in [Mestrano and Simpson 2011] referring to the two cases of Proposition 7.1 from there. Case (a) is when $H_1 \cap H_2 \cap H_3$ has dimension 0. It is a subscheme

[3] An algebraic argument is needed for the piece of P located at $R \cap U'$; letting A denote its coordinate algebra, u the equation of U', f the equation of U and g the equation of another plane through R, our hypothesis is $fuA = 0$ and the local piece of $P \cap (U' \cup R)$ corresponds to A/guA. Considering the exact sequence

$$A/guA \to A/(fA + gA) \oplus A/uA \to \mathbb{C} \to 0,$$

we see that if the required inequality $\ell(A/guA) \geq \ell(A/(fA + gA)) + \ell(fA)$ didn't hold we would have $guA = (fA + gA) \cap uA$ and $fA \cong A/uA$, hence also $uA \cong A/fA$. The exact sequence

$$0 \to guA \to uA \to A/(fA + gA)$$

becomes $0 \to g(A/fA) \to A/fA \xrightarrow{u} A/(fA + gA)$, which would give that multiplication by u on $A/(fA + gA)$ is injective, but that isn't possible since A has finite length.

of length 8 so we get

$$P = H_1 \cap H_2 \cap H_3.$$

A general such subscheme satisfies CB(2), and I. Dolgachev pointed out to us that these are called "Cayley octads". We shall treat the Cayley octads of case (a) secondly, since that will use one part of the discussion of case (b).

Case (b): This is when the subscheme $Y = H_1 \cap H_2 \cap H_3$ contains a pure 1-dimensional subscheme Y_1. Notice that Y_1 is a union of components of the curve[4] $H_1 \cap H_2$. On the other hand, by Lasker's theorem [Eisenbud et al. 1996, p. 314] if Y_1 were equal to $H_1 \cap H_2$ then there couldn't be a third quadric vanishing on Y_1. Therefore, Y_1 is a curve of degree ≤ 3.

We will now show that Y_1 doesn't contain a line. Suppose to the contrary that $R \subset Y_1$ is a line. Then all quadrics in our family contain R.

Choose a point p of P not on R, let U be the plane through R and p, and apply Lemma 3.1 with $n = 2$ to get $P \subset U \cup U'$. If $P \cap U'$ has length ≥ 5, it doesn't have four collinear points so it imposes 5 conditions on conics; hence we can apply Lemma 3.1 with $n = 1$ and get three residual points in a line, contradicting Lemma 3.2. Therefore $P \cap U$ has length ≥ 4, however since $P \cap R$ has length ≤ 2 by Lemma 3.2, the residual of $P \cap U$ with respect to R has length ≥ 2. Now, vanishing on R and on $P \cap U$ imposes 5 conditions on conics of U. Thus we may again apply Lemma 3.1 with $n = 1$ and get a residual consisting of 3 collinear points contradicting Lemma 3.2. This completes the proof that Y_1 does not contain a line.

That rules out almost all of the cases listed in [Mestrano and Simpson 2011, Lemma 7.4].

A next case is if Y_1 is a conic in a plane U. Then, it suffices to impose a single condition, $n = 1$ in Lemma 3.1, so $3 = 10 - c \geq n + 2$ and the residual subscheme consists of at least 3 points in a line. This contradicts Lemma 3.2, so Y_1 cannot be a plane conic.

The only remaining possibility for our curve of degree three, is that Y_1 could be a rational cubic curve not contained in a plane. It has to be a rational normal cubic, in particular smooth. The restriction of $\mathcal{O}_{\mathbb{P}^3}(2)$ to the rational curve has degree 6 so it has seven linearly independent sections; our three-dimensional family of quadrics is therefore the family of all quadrics passing through Y_1. They define Y_1 schematically, in particular $P \subset Y_1$.

This case will be of interest for our treatment of case (a) below. We have that P is a length 8 subscheme of the intersection $Y_1 \cap X$. For given Y_1 the space of choices of P is discrete, and as Y_1 moves any P becomes general. The family of such subschemes may therefore be identified with a covering of the space of choices

[4]Note that H_i cannot all vanish on some plane, otherwise by CB(1) for the residual P would have to be contained in the plane as we saw previously.

of rational normal cubic Y_1. The covering is determined, over a general point, by the choice of 8 out of the 15 points in $Y_1 \cap X$, or equivalently by the choice of the 7 complementary points.

The space of choices of Y_1 has dimension 12 (see [Mestrano and Simpson 2011, §6.2]). Therefore, this family cannot constitute an irreducible component of $\{P\}$. This completes the proof that case (b) cannot happen at a general point of an irreducible component.

Case (a): We start this discussion by continuing to look at the above 12-dimensional family of subschemes consisting of points in $X \cap Y_1$ for a smooth rational normal cubic curve Y_1.

We claim that the family of subschemes, and hence of bundles, obtained in this way is irreducible. This may be seen as follows. Any 6 points from $X \cap Y_1$ determine the rational normal cubic Y_1, so if we move a set of 6 points around to a different set, we get back to the same rational normal curve and this shows that the monodromy action includes permutations sending any subset of 6 points to any other one. On the other hand, there is a rational normal curve with first order tangency to X, and moving it a little bit induces a permutation of two points keeping the other points fixed. Therefore, the subgroup of the symmetric group contains a transposition. Now since it is 6-tuply transitive, it contains all the transpositions. Thus, the monodromy group is the full symmetric group and any group of 8 points can be moved to any other one. This shows that the family is irreducible.

As was pointed out at the end of Section 6.2 in [Mestrano and Simpson 2011], the space of obstructions at a general point in our family has dimension 1. The expected dimension is $4c_2 - 20 = 12$, so the Zariski tangent space to the moduli space has dimension 13; however, as noted above, any irreducible component has dimension ≥ 13 because of the existence of the extension. Therefore, a general point of our 12-dimensional family lies in a smooth open subset of a unique 13-dimensional irreducible component of the moduli space $\{E\}$ (notice here that the spaces $\{E, P\}$ and $\{P\}$ are also the same). As our 12-dimensional family is irreducible by the previous paragraph, this determines a canonical irreducible component of the moduli space.

This discussion corrects an error of notation in the second paragraph of the proof of Lemma 7.6 of [Mestrano and Simpson 2011], where it was stated that the irreducible 12-dimensional family of Cayley–Bacharach subschemes on the rational normal cubic was inside the type (a) subspace of the moduli space; but that family is clearly of type (b). Those phrases should be replaced by the argument of the previous paragraph showing that our 12-dimensional family is contained in a unique 13-dimensional irreducible component of the moduli space, whose general point is of type (a).

We now turn to consideration of the full set of irreducible components, whose general points are of type (a), that is to say bundles determined by Cayley octad subschemes P (since we showed in the previous part that type (b) cannot lead to a general point of a component).

The argument given in [Mestrano and Simpson 2011, §7.4], using the incidence variety suggested by A. Hirschowitz, shows that the existence of a canonically defined irreducible component implies irreducibility of the moduli space.

Let us recall here briefly how this works. We look at the full incidence scheme $\{X, P\}$ parametrizing smooth quintic hypersurfaces X together with l.c.i. subschemes $P \subset X$ of length 8 satisfying CB(2) of type (a). For a given $P \subset \mathbb{P}^3$ the space of quintics X containing it is a projective space and these all have the same dimension. So the fibration $\{X, P\} \to \{P\}$ is smooth, over the base that is an open subset in the Grassmannian Grass(3, 10) of 3-dimensional subspaces of $H^0(\mathcal{O}_{\mathbb{P}^3}(2))$. Thus, the full incidence variety $\{X, P\}$ is irreducible. There is a dense open subset of the space of quintics $\{X\}$, over which the sets of irreducible components of the fibers don't change locally. Thus, the fundamental group of this open set acts on the set of irreducible components of the fiber $\{P\}_X$ over a basepoint $X \in \{X\}$. This action is transitive, by irreducibility of the full incidence variety. On the other hand, we have described above a canonically defined irreducible component of $\{P\}_X$, containing the nearby generalizations of our 12-dimensional family of subschemes of a rational normal cubic curve. Since it is canonically defined, this component is preserved by the monodromy action. Transitivity now implies that $\{P\}_X$ has only a single irreducible component.

This completes the proof of irreducibility for $c_2 = 8$. The generic space of obstructions has dimension 1. That was seen for points on the rational normal cubic curve, at the end of §6.2 of [Mestrano and Simpson 2011]; however the moduli space has dimension 13, equal to the expected dimension plus 1, so the space of obstructions remains 1-dimensional at a general point.

As the dimension of the moduli space is equal to the expected dimension plus the dimension of the space of obstructions, we get that the moduli space is generically smooth, and in fact that was already the case at a point of the 12-dimensional family of subschemes on a rational normal cubic. Since $c = 7$ at a general point we have $h^1(E(1)) = 8 - 7 - 1 = 0$, to complete the corresponding column of our table.

For $c_2 = 9$. For the column $c_2 = 9$, see [Mestrano and Simpson 2011, Theorem 6.1 and Proposition 7.2], for the dimension 16 and general obstruction space of dimension 1. The proof of Proposition 7.2 there starts out by ruling out, for a general point of an irreducible component, all cases of Proposition 7.1 there except case (d), for which $c = 8$. Thus $h^1(E(1)) = 9 - 8 - 1 = 0$ for a general bundle, as we shall also see below.

We give here an alternate argument by dimension count to show that a general bundle in any irreducible component consists of a collection of 9 out of the 20 points on a degree 4 elliptic curve, intersection of two quadrics, intersected with X.

The expected dimension of $\{E\}$ is $4c_2 - 20 = 16$, and a general E determines a unique[5] extension hence a unique subscheme P of length 9. The dimension of any irreducible component of $\{E, P\}$ is ≥ 16 (notice that it coincides with the value of $3c_2 - 11$ too).

We first rule out the possibility that $c \leq 7$ for a general point. If there were a three-dimensional family of quadrics passing through P then they cannot intersect transversally in a zero-dimensional subscheme, since that would have length only 8 and so be unable to contain P. But if the intersection of the three quadrics has a component of positive dimension, then arguing much as in the previous section we can get a contradiction. Indeed, the space of length 9 subschemes contained in the intersection of X with two planes has dimension $\leq 3+3+9 = 15 < 16$, so any time Lemma 3.1 applies we immediately obtain a contradiction. The remaining case of points on a rational normal curve is ruled out by dimension.

We may therefore assume $c = 8$, from which it follows that any irreducible component of $\{P\}$ has dimension ≥ 16. It follows that a general P contains at least 7 points in general position on X. Let us explain the details of this argument, since this kind of dimension count has already been used several times above. Let $\boldsymbol{H} \subset \{P\}$ denote some component of the Hilbert scheme of subschemes we are interested in, that is to say l.c.i. subschemes $P \subset X$ of length 9 satisfying CB(2). Let

$$\boldsymbol{I} \subset \boldsymbol{H} \times X$$

be the incidence subscheme, whose fiber over a point $h \in \boldsymbol{H}$ is the subscheme P_s thereby parametrized. Suppose $p_1, \ldots, p_k$ is a collection of distinct points in X, and let $\boldsymbol{H}(p_1, \ldots, p_k) \subset \boldsymbol{H}$ be the closed subscheme parametrizing those P that contain $p_1, \ldots, p_k$. It may be inductively defined as follows: we have the incidence subvariety $\boldsymbol{I}(p_1, \ldots, p_k) \subset \boldsymbol{H}(p_1, \ldots, p_k) \times X$, and for a point p_{k+1} distinct from the other ones,

$$\boldsymbol{H}(p_1, \ldots, p_k, p_{k+1}) := \mathrm{pr}_2^{-1}(p_k) \subset \boldsymbol{I}(p_1, \ldots, p_k).$$

By induction we show that for general points p_i, $\boldsymbol{H}(p_1, \ldots, p_k)$ is nonempty of dimension $\geq 16-2k$ whenever $k \leq 7$. Assume it is known for $k-1$ but not true for k. That means that the map $\boldsymbol{I}(p_1, \ldots, p_{k-1}) \to X$ maps onto a closed subvariety; in other words, there is a curve $C \subset X$ depending on $p_1, \ldots, p_{k-1}$ and containing all of the subschemes parametrized by points of $\boldsymbol{H}(p_1, \ldots, p_{k-1})$. But then the space of

[5] An easy dimension count rules out the possibility that P be contained in a plane.

such subschemes has dimension $\leq 9-(k-1)$ (by [Briançon et al. 1981]), contradicting our inductive hypothesis since $9-(k-1) < 16-2(k-1)$ as $(k-1) < 16-9 = 7$.

After the 7 points in general position there remain two points. We may conclude that the dimension of a family of subschemes P, once the set theoretical locations of the points are known, is ≤ 2.

We now claim that if P is general, then for a general element H of our family of quadrics passing through P, the intersection $H \cap X$ is smooth. The proof is by a dimension count of the complementary family. If the $H \cap X$ is always singular, then the singular point is a basepoint (of the linear system on X), of which there are finitely many, so it is fixed. Thus, all the H are tangent to X at some point. The space of 2-dimensional linear systems tangent to $x \in X$ is a Grassmannian $\mathrm{Grass}(2, \mathbb{C}^7)$ of dimension 10. As the point moves in X we have a 12-dimensional space of choices of the linear system; and each one of these fixes the set-theoretical location of the points of P so by the previous paragraph, the corresponding space of P has dimension ≤ 2, so altogether we obtain that the family not satisfying our claimed condition has dimension ≤ 14. Since any component has dimension ≥ 16 it follows that the complementary family cannot constitute a component, which proves the claim.

Suppose $V := H^0(J_{P/\mathbb{P}^3}(2)) \subset \mathbb{C}^{10} = H^0(\mathcal{O}_{\mathbb{P}^3}(2))$ is the two-dimensional space of quadrics passing through our general point P. Then any deformation of the subspace $V \subset \mathbb{C}^{10}$ lifts to a deformation of P. This is because, by the previous claim, we can choose a general element of V corresponding to a quadric H_1 such that $H_1 \cap X$ is smooth. As the smooth curve deforms, our subscheme P of $(H_1 \cap X) \cap H_2$ becomes general since it is uniquely determined just by its multiplicities at each point.

From the above discussion it follows that a general point P in any irreducible component is obtained by choosing 9 out of the 20 points of $(H_1 \cap H_2) \cap X$ for a general pair of quadrics H_1, H_2. But since any 8 points determine the subspace $\langle H_1, H_2 \rangle$, the monodromy action on the set of 20 intersection points is 8-tuply transitive. By going around a curve $H_1 \cap H_2$ with a single simple tangent point to X, we get a transposition in the monodromy group; hence it contains all transpositions and it is the full symmetric group. Therefore, the set of choices of 9 points forms a single orbit under the monodromy group. This completes the proof that there is only one irreducible component of dimension 16.

The space of obstructions at a general point has dimension 1, see the discussion above Theorem 6.1 in [Mestrano and Simpson 2011]. This completes our review of the proof of Proposition 2.2.

For $c_2 \geq 10$. We will not be further reviewing the partial result of the case $c_2 = 10$ that was treated in [Mestrano and Simpson 2013], giving irreducibility of the open subset of the moduli space corresponding to seminatural cohomology as was stated in Proposition 2.4 above, since the argument is more involved and it is the subject of a distinct paper.

On the other hand, it will be useful to discuss in more detail the structure of $V(c_2)$.

Lemma 3.3. *For $c_2 \geq 11$, $V(c_2)$ is irreducible of dimension $3c_2 - 11$ and its general point corresponds to a set of points P in general position with respect to quadrics. The closure of $V(c_2)$ meets the boundary.*

Proof. See [Mestrano and Simpson 2011, Corollary 7.1], showing that for $c_2 \geq 11$, Σ_{c_2} contains an open dense subset $\Sigma_{c_2}^{10}$ consisting of collections P such that any colength 1 subscheme of P imposes vanishing of all quadrics. This is an open subset of the Hilbert scheme of all subschemes P of length c_2 so it is smooth, and it further contains an open dense subscheme where the points of P are distinct. The latter is an open subset of the symmetric product of X so it is irreducible.

The closure of $V(c_2)$ intersects the boundary, as was discussed in the proof of [Nijsse 1995, Proposition 3.2]. Indeed, choose a collection P_0 of distinct points that impose vanishing of quadrics but that doesn't satisfy $\mathrm{CB}(2)$. Deform this collection in a family P_t such that the general P_t (for $t \neq 0$) satisfies $\mathrm{CB}(2)$. Since all elements of the family impose the same number of conditions on quadrics, the space of Ext groups varies in a bundle with respect to the parameter t and we may choose a family of extensions such that the general one is locally free. But the special one is not locally free since P_0 didn't satisfy $\mathrm{CB}(2)$. This family gives a curve in $\Sigma_{c_2}^{10}$ with parameter $t \neq 0$, whose limiting sheaf at $t = 0$ is not locally free: we have a deformation to the boundary. □

Lemma 3.4. *For $c_2 = 10$, $V(10)$ is irreducible of dimension $3c_2 - 11 = 19$ and its general point corresponds to a subscheme P composed of 10 general points on a smooth intersection with a quadric $Y = X \cap H$. A general bundle in $V(10)$ has $h^1(E(1)) = 0$ so any deformation moving away from $V(10)$ will have seminatural cohomology, and only the irreducible component of $M(10)$ constructed in [Mestrano and Simpson 2013] contains $V(10)$.*

Proof. See [Nijsse 1995, Lemma 3.1]. General elements of any irreducible component correspond to subschemes P not contained in a plane, so the irreducible components of $V(10)$ correspond to those of Σ_{10} having the same dimension.

By [Mestrano and Simpson 2011, Corollary 7.1], Σ_{10} is pure of dimension 19. The stratum Σ_{10}^{8} consisting of extensions where P lies in the intersection of two quadrics, has dimension < 19. Indeed, the subscheme P is determined by the two-dimensional subspace of quadrics[6] and this has dimension 16, to which we should add 1 for the space of choices of extension: it comes out strictly less than 19.

[6] Unless they share a common plane but that case may also be dealt with by a dimension count: 3 for the choice of plane, plus 4 for the choice of line, plus at most 7 for the choice of points in the

Similarly, the dimension of the stratum Σ^7_{10} is strictly less than 19, and the strata Σ^c_{10} for $c \le 6$ may be ruled out using our previous line of argument with Lemma 3.1.

We conclude that the stratum Σ^9_{10} is dense in Σ_{10}. Here the extension class is determined (up to scaling) so $\{E, P\}$ and $\{P\}$ are the same, and $\{P\}$ is an open subset of the space $\{H, P\}$ parametrizing quadrics H together with $P \subset H \cap X$. The open subset is given by the conditions that no other quadrics vanish on P, and that P satisfies CB(2). But the space $\{H, P\}$ is irreducible.

Thus, $V(10)$ is irreducible and its general point parametrizes collections of 10 general points on a general smooth quadric section $Y = X \cap H$. One may now calculate with the standard exact sequence that for a general $E \in V(10)$, we have $h^1(E(1)) = 0$.

Recall by [Mestrano and Simpson 2013, Corollary 3.5] that the condition of having seminatural cohomology, for bundles in $M(10)$, is equivalent to the conjunction of two conditions[7] $h^1(E(1)) = 0$ and $h^0(E) = 0$. Bundles in $V(10)$ clearly don't satisfy the second condition because $V(10)$ is the locus where $h^0(E) > 0$. However, we have seen that a general point of $V(10)$ satisfies the first condition. On the other hand $V(10)$ is pure of dimension 19 whereas any component of $M(10)$ has dimension ≥ 20. Therefore, in any irreducible component of $M(10)$ containing $V(10)$, the general point has $h^0(E) = 0$, but also $h^1(E(1)) = 0$ since it is a generization of the general point of $V(10)$ that satisfies this condition. Therefore, any irreducible component of $M(10)$ containing V parametrizes, generically, bundles with seminatural cohomology.

It now follows from the main result of [Mestrano and Simpson 2013] (stated as Proposition 2.4 above) that any irreducible component of $M(10)$ containing $V(10)$ must be the unique component constructed there. □

4. The double dual stratification

Turn now to the proof of the main theorem on the moduli spaces for $c_2 \ge 10$. Our subsequent proofs will make use of O'Grady's [1993; 1996] techniques, as they were recalled and used by Nijsse [1995]. The main idea is to look at the boundary of the moduli spaces. His first main observation is the following:

Lemma 4.1 [O'Grady 1996, Proposition 3.3]. *The boundary of any irreducible component (or indeed, of any closed subset) of $M(c_2)$ has pure codimension* 1, *if it is nonempty.*

---

plane since they would otherwise all be in the plane and then we could ignore the choice of line, plus 1 for the choice of extension class, comes out to strictly less than 19.

[7] We use duality and Euler characteristic to rewrite the conditions of [Mestrano and Simpson 2013, Corollary 3.5].

The boundary is divided up into Uhlenbeck strata corresponding to the "number of delta-like singular instantons", which in the geometric picture corresponds to the number of points where the torsion-free sheaf is not a bundle, counted with correct multiplicities. A boundary stratum denoted $M(c_2, c_2 - d)$ parametrizes torsion-free sheaves F fitting into an exact sequence of the form

$$0 \to F \to E \xrightarrow{\sigma} S \to 0$$

where $E \in M(c_2 - d)$ is a stable locally free sheaf of degree 1 and $c_2(E) = c_2 - d$, and S is a finite coherent sheaf of length d so that $c_2(F) = c_2$. In this case $E = F^{**}$. We may think of $M(c_2, c_2 - d)$ as the moduli space of pairs (E, σ). Forgetting the quotient σ gives a smooth map

$$M(c_2, c_2 - d) \to M(c_2 - d),$$

sending F to its double dual. The fiber over E is the Grothendieck Quot scheme $\mathrm{Quot}(E, d)$ parametrizing quotients σ of E of length d.

Since we are dealing with sheaves of degree 1, all semistable points are stable and our objects have no nonscalar automorphisms. Hence the moduli spaces are fine, with a universal family existing étale-locally and well defined up to a scalar automorphism. We may view the double-dual map as being the relative Grothendieck Quot scheme of quotients of the universal object E^{univ} on $M(c_2 - d) \times X$ over $M(c_2 - d)$. Furthermore, locally on the Quot scheme the quotients are localized near a finite set of points, and we may trivialize the bundle E^{univ} near these points, so $M(c_2, c_2 - d)$ has a covering by, say, analytic open sets which are trivialized as products of open sets in the base $M(c_2 - d)$ with open sets in $\mathrm{Quot}(E, d)$ for any single choice of E. This is all to say that the map $M(c_2, c_2 - d) \to M(c_2 - d)$ may be viewed as a fibration in a fairly strong sense, with fiber $\mathrm{Quot}(E, d)$.

Li [1993, Proposition 6.4] shows that $\mathrm{Quot}(E, d)$ is irreducible with a dense open subset U parametrizing quotients which are given by a collection of d quotients of length 1 supported at distinct points of X:

Theorem 4.2 [Li 1993]. *Suppose E is a locally free sheaf of rank 2 on X. Then for any $d > 0$, $\mathrm{Quot}(E, d)$ is an irreducible scheme of dimension $3d$, containing a dense open subset parametrizing quotients $E \to S$ such that $S \cong \bigoplus \mathbb{C}_{y_i}$, where $\mathbb{C}_{y_i}$ is a skyscraper sheaf of length 1 supported at $y_i \in X$, and the y_i are distinct. This dense open set maps to $X^{(d)} - \mathrm{diag}$ (the space of choices of distinct d-tuple of points in X), with fiber over $\{y_i\}$ equal to $\prod_{i=1}^{d} \mathbb{P}(E_{y_i})$.*

Proof. See Proposition 6.4 in the appendix of [Li 1993]. Notice right away that U is an open subset of $\mathrm{Quot}(F, d)$, and that U fibers over the set $X^{(d)} - \mathrm{diag}$ of distinct d-tuples of points $(y_1, \ldots, y_d)$ (up to permutations). The fiber over a d-tuple $(y_1, \ldots, y_d)$ is the product of projective lines $\mathbb{P}(F_{y_i})$ of quotients of the vector

c_2	e.d.	$\dim(M)$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$
4	-4	2	–	–	–	–	–	–	–	–
5	0	3	5	–	–	–	–	–	–	–
6	4	7	6	8	–	–	–	–	–	–
7	8	9	10	9	11	–	–	–	–	–
8	12	13	12	13	12	14	–	–	–	–
9	16	16	16	15	16	15	17	–	–	–
10	20	20	19	19	18	19	18	20	–	–
11	24	24	23	22	22	21	22	21	23	–
12	28	28	27	26	25	25	24	25	24	26
≥ 13	$4c_2-20$	$4c_2-20$	$4c_2-21$	$\leq 4c_2-22$						

Table 2. Dimensions of strata.

spaces F_{y_i}. As $X^{(d)}$ – diag has dimension $2d$, and $\prod_{i=1}^{d} \mathbb{P}(F_{y_i})$ has dimension d, we get that U is a smooth open variety of dimension $3d$.

This theorem may also be viewed as a consequence of a more precise bound established by Ellingsrud and Lehn [1999], which will be stated as Theorem 7.6 below, needed for our arguments in Section 7. □

Corollary 4.3. *We have*

$$\dim(M(c_2; c_2')) = \dim(M(c_2')) + 3(c_2 - c_2').$$

If $M(c_2')$ is irreducible, then $M(c_2; c_2')$ and hence $\overline{M(c_2; c_2')}$ are irreducible.

Proof. The fibration $M(c_2; c_2') \to M(c_2')$ has fiber the Quot scheme whose dimension is $3(c_2 - c_2')$ by the previous proposition. Furthermore, these Quot schemes are irreducible so if the base is irreducible, so is the total space. □

Corollary 4.3 allows us to fill in the dimensions of the strata $M(c_2; c_2')$ in Table 2, starting from the dimensions of the moduli spaces given by Propositions 2.2 and 2.3. The entries in the second column are the expected dimension $4c_2 - 20$; in the third column the dimension of $M := M(c_2)$; and in the following columns, $\dim M(c_2, c_2 - d)$ for $d = 1, 2, \ldots$. The rule is to add 3 as you go diagonally down and to the right by one.

The first remark useful for interpreting this information is that any irreducible component of $\overline{M}(c_2)$ must have dimension at least equal to the expected dimension $4c_2 - 20$. In particular, a stratum with strictly smaller dimension must be a part of at least one irreducible component containing a bigger stratum. For $c_2 \geq 11$, we have

$$\dim(M(c_2, c_2')) < \dim(\overline{M}(c_2)) = 4c_2 - 20.$$

Hence, for $c_2 \geq 11$ the closures $\overline{M(c_2, c_2')}$ cannot themselves form irreducible components of $\overline{M}(c_2)$, in other words the irreducible components of $\overline{M}(c_2)$ are the same as those of $M(c_2)$. Notice, on the other hand, that $\bar{M}(10)$ contains two pieces of dimension 20, the locally free sheaves in $M(10)$ and the sheaves in $M(10,4)$ whose double duals come from $M(4)$.

Recall from Proposition 2.2 that the moduli spaces $M(c_2)$ are irreducible for $c_2 = 4, \ldots, 9$. It follows from Corollary 4.3 that the strata $M(c_2, c_2')$ are irreducible, for any $c_2' \leq 9$. In particular, the piece $\overline{M(10,4)}$ is irreducible, and its general point, representing a not locally free sheaf, is not confused with any point of $\overline{M(10)}$. Since the other strata of $\bar{M}(10)$ all have dimension < 20, it follows that $\overline{M(10,4)}$ is an irreducible component of $\bar{M}(10)$. One similarly gets from the table that $\bar{M}(c_2)$ has several irreducible components when $5 \leq c_2 \leq 9$.

5. Hartshorne's connectedness theorem

Hartshorne proves a connectedness theorem for locally complete intersections. Here is the version that we need.

Theorem 5.1 [Hartshorne 1962]. *Suppose Z is a locally complete intersection of dimension d. Then, any nonempty intersection of two irreducible components of Z has pure dimension $d-1$.*

Proof. See [Hartshorne 1962; Sawant 2011]. □

Corollary 5.2. *If the moduli space $\bar{M}$ is good and has two different irreducible components Z_1 and Z_2 meeting at a point z, then $Z_1 \cap Z_2$ has codimension 1 at z and the singular locus $\mathrm{Sing}(\bar{M})$ contains z and has codimension 1 at z.*

Proof. If $\bar{M}$ is good, then by Lemma 2.1 it is a locally complete intersection so Hartshorne's theorem applies: $Z_1 \cap Z_2$ has pure codimension 1. The intersection of two irreducible components is necessarily contained in the singular locus. □

We draw the following conclusions.

Corollary 5.3. *Suppose, for $c_2 \geq 10$, that two different irreducible components Z_1 and Z_2 of $\bar{M}$ meet at a point z. Then z is on the boundary.*

Proof. If z is not on the boundary, then by the previous corollary it is in a component of the singular locus having codimension 1 in M. We have seen in [Mestrano and Simpson 2011, Theorem 7.1] that for $c_2 \geq 10$, a piece of $\mathrm{Sing}(M)$ having codimension 1 in $M(c_2)$ has to be in $V(c_2)$, cf., Proposition 2.3 above. On the other hand $V(c_2)$ is irreducible, see Lemmas 3.3 and 3.4, so any such component of $\mathrm{Sing}(M)$ has to be equal to $V(c_2)$.

Recall that $\dim(V(c_2)) = 3c_2 - 11$ whereas the dimension of the moduli space is $4c_2 - 20$, thus for $c_2 \geq 11$ the singular locus has codimension ≥ 2, so the present situation could only occur for $c_2 = 10$.

But now by Lemma 3.4, $V(10)$ is contained in only one irreducible component of M, the one whose general point parametrizes bundles with seminatural cohomology. So, two distinct components cannot meet along $V(10)$. □

Next, recall one of Nijsse's theorems, connectedness of the moduli space.

Theorem 5.4 [Nijsse 1995]. *For $c_2 \geq 10$, the moduli space $\bar{M}(c_2)$ is connected.*

Proof. See [Nijsse 1995], Proposition 3.2. We have reviewed the argument in [Mestrano and Simpson 2016, Theorem 18.8]. □

Corollary 5.5. *Suppose Z is an irreducible component of $\bar{M}(c_2)$ for $c_2 \geq 10$. Then Z meets the boundary in a nonempty subset of codimension ≤ 1.*

Proof. The codimension 1 property is given by Lemma 4.1, so we just have to show that $\bar{Z}$ contains a boundary point.

For $c_2 \geq 10$, the first boundary stratum $M(c_2, c_2 - 1)$ has codimension 1, so it must meet at least one irreducible component of $\overline{M(c_2)}$, call it Z_0. Of course if $Z = Z_0$ we are done. Suppose $Z \subset M(c_2)$ is another irreducible component with $c_2 \geq 10$. By the connectedness of $\bar{M}(10)$, there exist a sequence of irreducible components $Z_0, \ldots, Z_k = \bar{Z}$ such that $Z_i \cap Z_{i+1}$ is nonempty. By Corollary 5.3, $Z_{k-1} \cap Z_k$ is contained in the boundary. □

6. Seminaturality along the 19-dimensional boundary strata

To treat the case $c_2 = 10$, we will apply the main result of our previous paper.

Proposition 6.1. *Suppose Z is an irreducible component of $M(10)$. Suppose that $\bar{Z}$ contains a point corresponding to a torsion-free sheaf F with $h^1(F(1)) = 0$. Then Z is the unique irreducible component containing the open set of bundles with seminatural cohomology, constructed in* [Mestrano and Simpson 2013].

Proof. The locus $V(c_2)$ of bundles with $h^0(E) \neq 0$ has dimension ≤ 19, so a general point of Z must have $h^0(E) = 0$. The hypothesis implies that a general point has $h^1(E(1)) = 0$. Thus, there is a nonempty dense open subset $Z' \subset Z$ parametrizing bundles with $h^0(E) = 0$ and $h^1(E(1)) = 0$. By [Mestrano and Simpson 2013, Corollary 3.5], these bundles have seminatural cohomology. Thus, our open set is $Z' = M(10)^{\mathrm{sn}}$, the moduli space of bundles with seminatural cohomology, shown to be irreducible in the main Theorem 0.2 of the same work recalled as Proposition 2.4 above. □

Using Proposition 6.1, and since we know by Corollary 5.5 that any irreducible component Z meets the boundary in a codimension 1 subset, in order to prove irreducibility of $M(10)$, it suffices to show that the torsion-free sheaves F parametrized by general points on the various irreducible components of the boundary of $\overline{M(10)}$ have $h^1(F(1)) = 0$.

The dimension is $\dim(Z) = 20$, so the boundary components will have dimension 19. Looking at the line $c_2 = 10$ in Table 2, we notice that there are three 19-dimensional boundary pieces, and a 20-dimensional piece which must constitute a different irreducible component. Consider first the 19-dimensional pieces,

$$M(10, 9),\ M(10, 8) \text{ and } M(10, 6).$$

Recall that $M(10, 10-d)$ consists generically of torsion-free sheaves F fitting into an exact sequence

$$0 \to F \to F^{**} \to S \to 0, \tag{6-1}$$

where F^{**} is a general point in the moduli space of stable bundles with $c_2 = 10-d$, and S is a general quotient of length d.

Proposition 6.2. *For a general point F in either of the three boundary pieces $M(10, 9)$, $M(10, 8)$ or $M(10, 6)$, we have $h^1(F(1)) = 0$.*

Proof. Notice that $\chi(F^{**}(1)) = 15 - c_2(F^{**}) \geq 6$ and by stability $h^2(F^{**}(1)) = h^0(F^{**}(-1)) = 0$, so $F^{**}(1)$ has at least six linearly independent sections. In particular, for a general quotient S of length 1, 2 or 4, consisting of the direct sum $S = \bigoplus S_x$ of general length 1 quotients $E_x \to S_x$ at 1, 2 or 4 distinct general points x, the map

$$H^0(F^{**}(1)) \to H^0(S)$$

will be surjective.

For a general point F^{**} in either $M(9)$, $M(8)$ or $M(6)$, we have $h^1(F^{**}(1)) = 0$. These results from [Mestrano and Simpson 2011] were recalled in Proposition 2.2, Table 1, and reviewed in Section 3. The long exact sequence associated to (6-1) now gives $h^1(F(1)) = 0$. □

This treats the 19-dimensional irreducible components of the boundary. There remains the piece $\overline{M(10, 4)}$ which has dimension 20. This is a separate irreducible component. It could meet $\overline{M(10)}$ along a 19-dimensional divisor, and we would like to show that $h^1(F(1)) = 0$ for the sheaves parametrized by this divisor. In particular, we are no longer in a completely generic situation so some further discussion is needed. This will be the topic of the next section.

7. The lowest stratum

The lowest stratum is $M(10, 4)$, which is therefore closed. We would like to understand the points in $\overline{M(10)} \cap M(10, 4)$. These are singular, so our main tool will be to look at where the singular locus of $\bar{M}(10)$ meets $M(10, 4)$. Denote this by

$$M(10, 4)^{\text{sing}} := \operatorname{Sing}(\bar{M}(10)) \cap M(10, 4).$$

In what follows, we give a somewhat explicit description of the lowest moduli space $M(4)$.

Lemma 7.1. *For $E \in M(4)$ we have $h^1(E)=0$, $h^0(E)=h^2(E)=3$, $h^0(E(1))=11$ and $h^1(E(1))=h^2(E(1))=0$.*

Proof. Choosing an element $s \in H^0(E)$ gives an exact sequence

$$0 \to \mathcal{O}_X \to E \to J_{P/X}(1) \to 0. \tag{7-1}$$

In [Mestrano and Simpson 2011] we have seen that $P \subset X \cap L$ is a subscheme of length 4 in the intersection of X with a line $L \subset \mathbb{P}^3$. As P spans L, the space of linear forms vanishing on P is the same as the space of linear forms vanishing on L, so $H^0(J_{P/X}(1)) \cong \mathbb{C}^2$. In the long exact sequence associated to (7-1), note that $H^1(\mathcal{O}_X)=0$, giving

$$0 \to H^0(\mathcal{O}_X) \to H^0(E) \to H^0(J_{P/X}(1)) \to 0,$$

hence $H^0(E) \cong \mathbb{C}^3$. By duality, $H^2(E) \cong \mathbb{C}^3$, and the Euler characteristic of E is 6, so $H^1(E)=0$.

For $E(1)$, note that $H^2(E(1))=0$ by stability and duality, and (7-1) gives an exact sequence

$$0 \to H^1(E(1)) \to H^1(J_{P/X}(2)) \to H^2(\mathcal{O}_X(1)) \to 0.$$

On the other hand, $H^1(J_{P/X}(2)) \cong \mathbb{C}$ corresponding to the length 4 of P, minus the dimension 3 of the space of sections of $\mathcal{O}_P(2)$ coming from global quadrics (since the space of quadrics on L has dimension 3). This gives $H^1(E(1))=0$. The Euler characteristic then gives $h^0(E(1))=11$. This is also seen in the first part of the exact sequence, where $H^0(\mathcal{O}_X(1))=\mathbb{C}^4$ and $H^0(J_{P/X}(2)) \cong \mathbb{C}^7$. □

If $p \in \mathbb{P}^3$, let $G \cong \mathbb{C}^3$ be the space of linear generators of the ideal of p, that is to say $G := H^0(J_{p/\mathbb{P}^3}(1))$, and consider the natural exact sequence of sheaves on $\mathbb{P}^3$

$$0 \to \mathcal{O}_{\mathbb{P}^3}(-1) \to \mathcal{O}_{\mathbb{P}^3} \otimes G^* \to \mathcal{R}_p \to 0.$$

Here the cokernel sheaf $\mathcal{R}_p$ is a reflexive sheaf of degree 1, and $c_2(\mathcal{R}_p)$ is the class of a line. The restriction $\mathcal{R}_p|_X$ therefore has $c_2=5$. If $p \in X$, it is torsion-free but not locally free, giving a point in $M(5,4)$. It turns out that these sheaves account for all of $M(4)$ and $M(5)$.

Theorem 7.2. *Suppose $E \in M(4)$. Then there is a unique point $p \in X$ such that E is generated by global sections outside of p, and $\mathcal{R}_p|_X$ is isomorphic to the subsheaf of E generated by global sections. This fits into an exact sequence*

$$0 \to \mathcal{R}_p|_X \to E \to S \to 0,$$

where S has length 1, *in particular* $E \cong (\mathcal{R}_p|_X)^{**}$. *The correspondence* $E \leftrightarrow p$ *establishes an isomorphism* $M(4) \cong X$.

For $E' \in M(5)$, *there exists a unique point* $p \in \mathbb{P}^3 - X$ *such that* $E' \cong \mathcal{R}_p|_X$. *This correspondence establishes an isomorphism* $\overline{M(5)} \cong \mathbb{P}^3$ *such that the boundary component* $M(5,4) \cap \overline{M(5)}$ *is exactly* $X \subset \mathbb{P}^3$. *Note however that* $M(5,4)$ *itself has dimension strictly bigger than* 3 *and constitutes another irreducible component of* $\overline{M}(5)$.

Proof. Consider the exact sequence (7-1). The space $H^0(J_{P/X}(1))$ consists of linear forms on X (or equivalently, on $\mathbb{P}^3$), which vanish along P. However, a linear form which vanishes on P also vanishes on L. In particular, elements of $H^0(J_{P/X}(1))$ generate $J_{X\cap L/X}(1)$, which has colength 1 in $J_{P/X}(1)$.

Let $R \subset E$ be the subsheaf generated by global sections, and let S be the cokernel in the exact sequence

$$0 \to R \to E \to S \to 0.$$

We also have the exact sequence

$$0 \to J_{X\cap L/X}(1) \to J_{P/X}(1) \to S \to 0$$

so S has length 1. It is supported on a point p. The sheaf R is generated by three global sections so we have an exact sequence

$$0 \to \ker \to \mathcal{O}_X^3 \to R \to 0.$$

The kernel is a saturated subsheaf, hence locally free, and by looking at its degree we have $\ker = \mathcal{O}_X(-1)$. Thus, R is the cokernel of a map $\mathcal{O}_X(-1) \to \mathcal{O}_X^3$ given by three linear forms; these linear forms are a basis for the space of forms vanishing at the point p. We see that R is the restriction to X of the sheaf $\mathcal{R}_p$ described above, hence $E \cong (\mathcal{R}_p|_X)^{**}$. The map $E \mapsto p$ gives a map $M(4) \to X$, with inverse $p \mapsto (\mathcal{R}_p|_X)^{**}$.

The second paragraph, about $\overline{M(5)}$, is not actually needed later and we leave it to the reader. □

Even though the moduli space $M(4)$ is smooth, it has much more than the expected dimension, and the space of coobstructions is nontrivial. It will be useful to understand the coobstructions, because if $F \in M(10,4)$ is a torsion-free sheaf with $F^{**} = E$ then coobstructions for F come from coobstructions for E which preserve the subsheaf $F \subset E$.

Lemma 7.3. *Suppose* $E \in M(4)$. *A general coobstruction* $\phi : E \to E(1)$ *has generically distinct eigenvalues with an irreducible spectral variety in* $\mathrm{Tot}(K_X)$.

Proof. It suffices to write down a map $\phi : E \to E(1)$ with generically distinct eigenvalues and irreducible spectral variety. To do this, we construct a map ϕ_R :

$R \to R(1)$ using the expression $R = \mathcal{R}_p|_X$. The exact sequence defining $\mathcal{R}_p$ extends to the Koszul resolution, a long exact sequence

$$0 \to \mathcal{O}_{\mathbb{P}^3}(-1) \to \mathcal{O}_{\mathbb{P}^3}^3 \to \mathcal{O}_{\mathbb{P}^3}(1)^3 \to J_{p/\mathbb{P}^3}(2) \to 0.$$

Thus $\mathcal{R}_p$ may be viewed as the image of the middle map. Without loss of generality, p is the origin in an affine system of coordinates (x, y, z) for $\mathbb{A}^3 \subset \mathbb{P}^3$, and the coordinate functions are the three coefficients of the maps on the left and right in the Koszul sequence. The 3×3 matrix in the middle is

$$K := \begin{pmatrix} 0 & z & -y \\ -z & 0 & x \\ y & -x & 0 \end{pmatrix}.$$

Any 3×3 matrix of constants Φ gives a composed map

$$\phi_R : \mathcal{R}_p \hookrightarrow \mathcal{O}_{\mathbb{P}^3}(1)^3 \xrightarrow{\Phi} \mathcal{O}_{\mathbb{P}^3}(1)^3 \to \mathcal{R}_p(1).$$

Use the first two columns of K to give a map $k : \mathcal{O}_{\mathbb{P}^3}^2 \to \mathcal{R}_p$ which is an isomorphism over an open set. On the other hand, the projection onto the first two coordinates gives a map $q : \mathcal{R}_p \to \mathcal{O}_{\mathbb{P}^3}(1)^2$ which is, again, an isomorphism over an open set. The composition of these two is the map given by the upper 2×2 square of K,

$$qk = K_{2,2} := \begin{pmatrix} 0 & z \\ -z & 0 \end{pmatrix}.$$

We can now analyze the map ϕ_R by noting that $q\phi_R k = K_{2,3}\Phi K_{3,2}$ where $K_{2,3}$ and $K_{3,2}$ are respectively the upper 2×3 and left 3×2 blocks of K. Over the open set where q and k are isomorphisms,

$$q\phi_R q^{-1} = q\phi_R k(qk)^{-1} = K_{2,3}\Phi K_{3,2}K_{2,2}^{-1}.$$

Now

$$K_{3,2}K_{2,2}^{-1} = \begin{pmatrix} 0 & z \\ -z & 0 \\ y & -x \end{pmatrix} \cdot \begin{pmatrix} 0 & -1/z \\ 1/z & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -x/z & -y/z \end{pmatrix}.$$

Suppose

$$\Phi = \begin{pmatrix} \alpha & \beta & \gamma \\ \delta & \epsilon & \psi \\ \chi & \theta & \rho \end{pmatrix}.$$

Then

$$
\begin{aligned}
q\phi_R q^{-1} &= K_{2,3}\Phi K_{3,2}K_{2,2}^{-1}\\
&= \begin{pmatrix} 0 & z & -y\\ -z & 0 & x\end{pmatrix}\cdot\begin{pmatrix}\alpha & \beta & \gamma\\ \delta & \epsilon & \psi\\ \chi & \theta & \rho\end{pmatrix}\cdot\begin{pmatrix}1 & 0\\ 0 & 1\\ -x/z & -y/z\end{pmatrix}\\
&= \begin{pmatrix} 0 & z & -y\\ -z & 0 & x\end{pmatrix}\cdot\begin{pmatrix}\alpha-\gamma x/z & \beta-\gamma y/z\\ \delta-\psi x/z & \epsilon-\psi y/z\\ \chi-\rho x/z & \theta-\rho y/z\end{pmatrix}\\
&= \begin{pmatrix} z\delta-\psi x-y\chi+\rho xy/z & z\epsilon-\psi y+y\theta-\rho y^2/z\\ -z\alpha+\gamma x-x\chi+\rho x^2/z & -z\beta+\gamma y+x\theta-\rho xy/z\end{pmatrix}.
\end{aligned}
$$

Notice that the trace of this matrix is

$$\mathrm{Tr}(\phi)=x(\theta-\psi)+y(\gamma-\chi)+z(\delta-\beta),$$

which is a section of $H^0(\mathcal{O}_{\mathbb{P}^3}(1))$ vanishing at p. A coobstruction should have trace zero, so we should impose three linear conditions

$$\theta=\psi,\quad \chi=\gamma,\quad \delta=\beta,$$

which together just say that Φ is a symmetric matrix. Our expression simplifies to

$$q\phi_R q^{-1}=\begin{pmatrix}\beta z-\psi x-\gamma y+\rho xy/z & \epsilon z-\rho y^2/z\\ -\alpha z+\rho x^2/z & -\beta z+\psi x+\gamma y-\rho xy/z\end{pmatrix}.$$

Now, restrict $\mathcal{R}_p$ to X to get the sheaf R, take its double dual to get $E=R^{**}$, and consider the induced map $\phi: E\to E(1)$. Over the intersection of our open set with X, this will have the same formula. We can furthermore restrict to the curve $Y\subset X$ given by the intersection with the plane $y=0$. Note that X is in general position subject to the condition that it contains the point p. Setting $y=0$ the above matrix becomes

$$(q\phi q^{-1})|_{y=0}=\begin{pmatrix}\beta z-\psi x & \epsilon z\\ -\alpha z+\rho x^2/z & -\beta z+\psi x\end{pmatrix}.$$

Choose for example $\beta=\psi=0$ and $\alpha=\rho=\epsilon=1$, giving the matrix whose determinant is

$$\det\begin{pmatrix}0 & z\\ x^2/z-z & 0\end{pmatrix}=z^2-x^2=(z+x)(z-x).$$

The eigenvalues of $\phi|_Y$ are therefore $\pm\sqrt{(z+x)(z-x)}$, generically distinct. For a general choice of the surface X, our curve $Y=X\cap(y=0)$ will intersect the planes $x=z$ and $x=-z$ transversally, so the two eigenvalues of $\phi|_Y$ are permuted when going around points in the ramification locus different from p. This provides an explicit example of ϕ for which the spectral variety is irreducible, completing the proof of the lemma. We included the detailed calculations because they look to be useful if one wants to write down explicitly the spectral varieties. □

Turn now to the study of the boundary component $M(10, 4)$ consisting of torsion-free sheaves in $\overline{M}(10)$ which come from bundles in $M(4)$. A point in $M(10, 4)$ consists of a torsion-free sheaf F in an exact sequence of the form (6-1)

$$0 \to F \to E \xrightarrow{\sigma} S \to 0,$$

where $E = F^{**}$ is a point in $M(4)$, and S is a length 6 quotient.

The basic description of the space of obstructions as dual to the space of K_X-twisted endomorphisms still holds for torsion-free sheaves. Thus, the obstruction space for F is $\mathrm{Hom}^o(F, F(1))^*$. A coobstruction is a map $\phi : F \to F(1) = F \otimes K_X$ with $\mathrm{Tr}(\phi) = 0$, which is a kind of *Higgs field*. Since the moduli space is good, a point F is in $\mathrm{Sing}(\overline{M}(10))$ if and only if the obstruction space is nonzero, that is to say, if and only if there exists a nonzero trace-free $\phi : F \to F(1)$.

To give a map ϕ is the same thing as to give a map $\varphi : E \to E(1)$ compatible with the quotient map $E \to S$, in other words fitting into a commutative square with σ, for an induced map $\varphi_S : S \to S$. The maps φ, coobstructions for E, were studied in Lemma 7.3 above.

Let $\mathbb{P}(E) \to X$ denote the Grothendieck projective space bundle. A point in $\mathbb{P}(E)$ is a pair (x, s), where $x \in X$ and $s : E_x \to S_x$ is a length one quotient of the fiber. Suppose we are given a map $\varphi : E \to E(1)$. We can consider the *internal spectral variety*

$$\mathrm{Sp}_E(\varphi) \subset \mathbb{P}(E),$$

defined as the set of points $(x, s) \in \mathbb{P}(E)$ such that there is a commutative diagram

$$\begin{array}{ccc} E_x & \xrightarrow{\varphi(x)} & E_x \\ \downarrow & & \downarrow \\ S_x & \longrightarrow & S_x. \end{array}$$

The term “internal” signifies that it is a subvariety of $\mathbb{P}(E)$ as opposed to the classical spectral variety which is a subvariety of the total space of K_X. Here, we have only given $\mathrm{Sp}_E(\varphi)$ a structure of closed subset of $\mathbb{P}(E)$, hence of reduced subvariety. It would be interesting to give it an appropriate scheme structure which could be nonreduced in case φ is nilpotent, but that will not be needed here.

Corollary 7.4. *Suppose $E \in M(4)$ and $\varphi : E \to E(1)$ is a general coobstruction. Then the internal spectral variety $\mathrm{Sp}_E(\varphi)$ has a single irreducible component of dimension 2. A quotient $E \to S$ consisting of a disjoint sum of length one quotients $s_i : E_{x_i} \to S_i$ with $S = \bigoplus S_i$ and the points x_i disjoint, is compatible with φ if and only if the points $(x_i, s_i) \in \mathbb{P}(E)$ lie on the internal spectral variety $\mathrm{Sp}_E(\varphi)$.*

Proof. Notice if $z \in X$ is a point such that $\varphi(z) = 0$, then the whole fiber $\mathbb{P}(E)_z \cong \mathbb{P}^1$ is in $\mathrm{Sp}_E(\varphi)$. In particular, if such a point exists then the map $\mathrm{Sp}_E(\varphi) \to X$ will not be finite.

A first remark is that the zero-set of φ is 0-dimensional. Indeed, if φ vanished along a divisor D, then $D \in |\mathcal{O}_X(n)|$ for $n \geq 1$ and $\varphi : F \to F(1-n)$. This is possible only if $n = 1$ and $\varphi : F \to F$ is a scalar endomorphism (since F is stable). However, the trace of the coobstruction vanishes, so the scalar φ would have to be zero, which we are assuming is not the case.

At an isolated point z with $\varphi(z) = 0$, the fiber of the projection $\mathrm{Sp}_E(\varphi) \to X$ contains the whole $\mathbb{P}(E_z) = \mathbb{P}^1$. However, these can contribute at most irreducible components of dimension ≤ 1 (although we conjecture that in fact these fibers are contained in the closure of the 2-dimensional component so that $\mathrm{Sp}_E(\varphi)$ is irreducible).

Away from such fibers, the internal spectral variety is isomorphic to the external one, a two-sheeted covering of X, and by Lemma 7.3, for a general φ the monodromy of this covering interchanges the sheets, so it is irreducible. Thus, $\mathrm{Sp}_E(\varphi)$ has a single irreducible component of dimension 2, and it maps to X by a generically finite (2 to 1) map.

The second statement, that a quotient consisting of a direct sum of length one quotients, is compatible with φ if and only if the corresponding points lie on $\mathrm{Sp}_E(\varphi)$, is immediate from the definition. □

Definition 7.5. A triple (E, φ, σ) where $E \in M(4)$, $\varphi : E \to E(1)$ is a nonnilpotent map, and $\sigma = \bigoplus s_x$ is a quotient composed of six length 1 quotients over distinct points, compatible with φ as in the previous Corollary 7.4, leads to an obstructed point $F = F_{(E,\varphi,\sigma)} \in M(10, 4)^{\mathrm{sing}}$ obtained by setting $F := \ker(\sigma)$. Such a point will be called *usual*.

Ellingsrud and Lehn have given a very nice description of the Grothendieck quotient scheme of a bundle of rank r on a smooth surface. It extends the basic idea of Li's theorem which we already stated as Theorem 4.2 above, and will allow us to count dimensions of strata in $M(10, 4)$.

Theorem 7.6 [Ellingsrud and Lehn 1999]. *The quotient scheme parametrizing quotients of a locally free sheaf $\mathcal{O}_X^r$ of rank r on a smooth surface X, located at a given point $x \in X$, and of length ℓ, is irreducible of dimension $r\ell - 1$.*

Proof. See [Ellingsrud and Lehn 1999]. We have given the local version of the statement here. □

In our case, $r = 2$ so the dimension of the local quotient scheme is $2\ell - 1$.

A given quotient $E \to S$ decomposes as a direct sum of quotients $E \to S_i$ located at distinct points $x_i \in X$. Order these by decreasing length, and define the *length vector* of S to be the sequence $(\ell_1, \ldots, \ell_k)$ of lengths $\ell_i = \ell(S_i)$ with $\ell_i \geq \ell_{i+1}$.

This leads to a stratification of the Quot scheme into strata labeled by length vectors. By [Ellingsrud and Lehn 1999], the dimension of the space of quotients supported at a single (but not fixed) point x_i and having length ℓ_i, is $2\ell_i + 1$, giving the following dimension count.

Corollary 7.7. *For a fixed bundle E of rank 2, the dimension of the stratum associated to length vector $(\ell_1, \dots, \ell_k)$ in the* Quot-*scheme of quotients $E \to S$ with total length $\ell = \sum_{i=1}^{k} \ell_i$, is*

$$\sum (2\ell_i + 1) = 2\ell + k.$$

Recall that the moduli space $M(4)$ has dimension 2, so the dimension of the stratum of $M(10, 4)$ corresponding to a vector $(\ell_1, \dots, \ell_k)$ is $14 + k$. In particular, $M(10, 4)$ has a single stratum $(1, 1, 1, 1, 1, 1)$ of dimension 20, corresponding to quotients which are direct sums of length one quotients supported at distinct points, and a single stratum $(2, 1, 1, 1, 1)$ of length 19. This yields the following corollary.

Corollary 7.8. *If $Z' \subset M(10, 4)$ is any 19-dimensional irreducible subvariety, then either Z' is equal to the stratum $(2,1,1,1,1)$, or else the general point on Z' consists of a direct sum of six length 1 quotients supported over six distinct points of X.*

Proposition 7.9. *The singular locus $M(10, 4)^{\text{sing}}$ has only one irreducible component of dimension 19. This irreducible component has a nonempty dense open subset consisting of the usual points (Definition 7.5). For a usual point, the coobstruction φ is unique up to a scalar, so this open set may be viewed as the moduli space of usual triples (E, φ, σ), which is irreducible.*

Proof. Suppose $Z' \subset M(10, 4)^{\text{sing}}$ is an irreducible component. Consider the two cases given by Corollary 7.8.

(i) If Z' contains an open set consisting of points which are direct sums of six length 1 quotients supported on distinct points of X, then this open set parametrizes usual triples. Furthermore, a point in this open set corresponds to a choice of (E, φ) together with six points on the internal spectral variety $\mathrm{Sp}_E(\varphi)$. We count the dimension of this piece as follows.

Let $M'(4)$ denote the moduli space of pairs (E, φ) with $E \in M(4)$ and φ a nonzero coobstruction for E. The space of coobstructions for any $E \in M(4)$, has dimension 6 and the family of these spaces forms a vector bundle over $M(4)$ (more precisely, a twisted vector bundle twisted by the obstruction class for existence of a universal family over $M(4)$). Thus, the moduli space of pairs has a fibration $M'(4) \to M(4)$ whose fibers are $\mathbb{P}^5$. In particular, $M'(4)$ is a smooth irreducible variety of dimension 7.

For a general such (E, φ) the moduli space of usual triples has dimension ≤ 12, with a unique 12-dimensional piece corresponding to a general choice of 6 points on the unique 2-dimensional irreducible component of $\mathrm{Sp}_E(\varphi)$. This gives the 19-dimensional component of $M(10, 4)^{\text{sing}}$ mentioned in the proposition.

Suppose (E,φ) is not general, that is to say, contained in some subvariety of $M'(4)$ of dimension ≤ 6. Then, as φ is nonzero, even though we no longer can say that it is irreducible, in any case the internal spectral variety $\mathrm{Sp}_E(\varphi)$ has dimension 2 so the space of choices of 6 general points on it has dimension ≤ 12, and this contributes at most subvarieties of dimension ≤ 18 in $M(10,4)^{\mathrm{sing}}$. This shows that in the first case (i) of Corollary 7.8, we obtain the conclusion of the proposition.

(ii) Suppose Z' is equal to the stratum of $M(10,4)$ corresponding to length vector $(2,1,1,1,1)$. In this case, we show that a general point of Z' has no nonzero coobstructions, contradicting the hypothesis that $Z' \subset M(10,4)^{\mathrm{sing}}$ and showing that this case cannot occur.

Fix $E \in M(4)$. The space of coobstructions of E has dimension 6. Suppose $E \to S_1$ is a quotient of length 2. If it is just the whole fiber of E over x_1, then it is automatically compatible with any coobstruction. However, these quotients contribute only a 2-dimensional subspace of the space of such quotients which has dimension 5 by [Ellingsrud and Lehn 1999]. Thus, these points don't contribute general points. On the other hand, a general quotient of length 2 corresponds to an infinitesimal tangent vector in $\mathbb{P}(E)$, and the condition that this vector be contained in $\mathrm{Sp}_E(\varphi)$ imposes two conditions on φ. Therefore, the space of coobstructions compatible with S_1 has dimension ≤ 4. Next, given a nonzero coobstruction in that subspace, a general quotient $E \to S_2$ of length 1 will not be compatible, so imposing compatibility with S_1 and S_2 leads to a space of coobstructions of dimension ≤ 3. Continuing in this way, we see that imposing the condition of compatibility of φ with a general quotient $S = S_1 \oplus \cdots \oplus S_5$ in the stratum $(2,1,1,1,1)$ leads to $\varphi = 0$. Thus, a general point of this stratum has no nonzero coobstructions as we have claimed, and this case (ii) cannot occur.

Hence, the only case from Corollary 7.8 which can contribute a 19-dimensional stratum, contributes the single irreducible component described in the statement of the proposition. One may note that φ is uniquely determined for a general set of six points on its internal spectral variety, since the first 5 points are general in $\mathbb{P}(E)$ and impose linearly independent conditions. □

Corollary 7.10. *Suppose $M(10,4) \cap \overline{M(10)}$ is nonempty. Then it is the unique 19-dimensional irreducible component of usual triples in $M(10,4)^{\mathrm{sing}}$ identified by Proposition 7.9.*

Proof. By Hartshorne's theorem, the intersection $M(10,4) \cap \overline{M(10)}$ has pure dimension 19 if it is nonempty. This could also be seen from O'Grady's lemma that the boundary of $\overline{M(10)}$ has pure dimension 19. However, any point in this intersection is singular. By Proposition 7.9, the singular locus $M(10,4)^{\mathrm{sing}}$ has only one irreducible component of dimension 19, and it is the closure of the space of usual triples. □

If the intersection $M(10,4) \cap \overline{M(10)}$ is nonempty, the torsion-free sheaves F parametrized by general points satisfy $h^1(F(1)) = 0$. We show this by a dimension estimate using [Ellingsrud and Lehn 1999]. The more precise information about $M(10,4)^{\text{sing}}$ given in Proposition 7.9, while not really needed for the proof at $c_2 = 10$, will be useful in treating the case of $c_2 = 11$ in Section 9.

Proposition 7.11. *The subspace of $M(10,4)$ consisting of points F such that $h^1(F(1)) \geq 1$ has codimension ≥ 2.*

Proof. Use the exact sequence $0 \to F \to E \to S \to 0$, where $E \in M(4)$. One has $h^1(E(1)) = 0$ for all $E \in M(4)$, see Lemma 7.1. Therefore, $h^1(F(1)) = 0$ is equivalent to saying that the map

$$H^0(E(1)) \to H^0(S(1)) \cong \mathbb{C}^6 \tag{7-2}$$

is surjective.

Considering the theorem of Ellingsrud and Lehn [1999], there are two strata to be looked at: the case of a direct sum of six quotients of length 1 over distinct points, to be treated below; and the case of a direct sum of four quotients of length 1 and one quotient of length 2. However, this latter stratum already has codimension 1, and it is irreducible. So, for this stratum it suffices to note that a general quotient $E \to S$ in it leads to a surjective map (7-2), which may be seen by a classical general position argument, placing first the quotient of length 2.

Consider now the stratum of quotients which are the direct sum of six length 1 quotients s_i at distinct points $x_i \in X$. Fix the bundle E. The space of choices of the six quotients (x_i, s_i) has dimension 18. We claim that the space of choices such that (7-2) is not surjective, has codimension ≥ 2.

Note that $h^0(E(1)) = 11$. Given six quotients (x_i, s_i), if the map (7-2) (with $S = \bigoplus S_i$) is not surjective, then its kernel has dimension ≥ 6, so if we choose five additional points $(y_j, t_j) \in \mathbb{P}(E)$ with $t_j : E_{y_j} \to T_j$ for T_i of length 1, the total evaluation map

$$H^0(E(1)) \to \bigoplus_{i=1}^{6} S_i(1) \oplus \bigoplus_{j=1}^{5} T_j(1) \tag{7-3}$$

has a nontrivial kernel. Consider the variety

$$W := \{(u, \dots, (x_i, s_i), \dots, \dots, (y_j, t_j), \dots) \mid 0 \neq u \in H^0(E(1)), s_i(u) = 0, t_j(u) = 0\}$$

with the nonzero section u taken up to multiplication by a scalar.

Let $Q_6'(E)$ and $Q_5'(E)$ denote the open subsets of the quotient schemes of length 6 and length 5 quotients of E respectively, open subsets consisting of quotients which are direct sums of length one quotients over distinct points. Let $K \subset Q_6'(E)$ denote the locus of quotients $E \to S$ such that the kernel sheaf F has $h^1(F(1)) \geq 1$.

It is a proper closed subset, since it is easy to see that a general quotient $E \to S$ leads to a surjection (7-2). The above argument with (7-3) shows that $K \times Q'_5(E) \subset p(W)$, where $p : W \to Q'_6(E) \times Q'_5(E)$ is the projection forgetting the first variable u. Our goal is to show that K has dimension ≤ 16.

We claim that W has dimension ≤ 32 and has a single irreducible component of dimension 32. To see this, start by noting that the choice of u lies in the projective space $\mathbb{P}^{10}$ associated to $H^0(E(1)) \cong \mathbb{C}^{11}$.

For a section u which is special in the sense that its scheme of zeros has positive dimension, the locus of choices of (x_i, s_i) and (y_j, t_j) has dimension ≤ 22, but might have several irreducible components depending on whether the points are on the zero-set of u or not. However, the space of sections u which are special in this sense is equal to the space of pairs $u' \in H^0(E)$, $u'' \in H^0(\mathcal{O}_X(1))$ up to scalars for both pieces, and this has dimension $2+3=5$, which is much smaller than the dimension of the space of all sections u. Therefore, these pieces don't contribute anything of dimension higher than 27.

For a section u which is not special in the sense of the previous paragraph, the space of choices of a single length 1 quotient (x, s) which vanishes on the section, has a single irreducible component of dimension 2. It might possibly have some pieces of dimension 1 corresponding to quotients located at the zeros of u (although we don't think so). Hence, the space of choices of point in W lying over the section u has dimension ≤ 22 and has a single irreducible component of dimension 22.

Putting these together over $\mathbb{P}^{10}$, the dimension of W is ≤ 32 and it has a single irreducible component of dimension 32, as claimed. Its image $p(W)$ therefore also has dimension ≤ 32, and has at most one irreducible component of dimension 32. Denote this component, if it exists, by $p(W)'$.

Suppose now that K had an irreducible component K' of dimension 17. Then $K' \times Q'_5(E) \subset p(W)$, but $\dim(Q'_5(E)) = 15$ so $p(W)'$ would exist and would be equal to $K' \times Q'_5(E)$. However, $p(W)'$ is symmetric under permutation of the 11 different variables (x, s) and (y, t), but that would then imply that $P(W)'$ was the whole of $Q'_6(E) \times Q'_5(E)$ which is not the case. Therefore, K must have codimension ≥ 2. This completes the proof of the proposition. □

Corollary 7.12. *Suppose $M(10,4) \cap \overline{M(10)}$ is nonempty. Then a general point of this intersection corresponds to a torsion-free sheaf with $h^1(F(1)) = 0$.*

Proof. By Hartshorne's or O'Grady's theorem, if the intersection is nonempty then it has pure dimension 19. However, the space of torsion-free sheaves $F \in M(10,4)$ with $h^1(F(1)) > 0$ has dimension ≤ 18 by Proposition 7.11. Thus, a general point in any irreducible component of $M(10,4) \cap \overline{M(10)}$ must have $h^1(F(1)) = 0$. In fact there can be at most one irreducible component, by Corollary 7.10. □

8. Irreducibility for $c_2 = 10$

Corollary 8.1. *Suppose Z is an irreducible component of $M(10)$. Then, for a general point F in any irreducible component of the intersection of $\bar{Z}$ with the boundary, we have $h^1(F(1)) = 0$.*

Proof. By O'Grady's lemma, the intersection of $\bar{Z}$ with the boundary has pure dimension 19. By considering the line $c_2 = 10$ in the Table 2, this subset must be a union of some of the irreducible subsets $\overline{M(10, 9)}$, $\overline{M(10, 8)}$, $\overline{M(10, 6)}$, and the unique 19-dimensional irreducible component of $M(10, 4)^{\text{sing}}$ given by Proposition 7.9. Combining Proposition 6.2 and Corollary 7.12, we conclude that any one of these irreducible components of the intersection of $\bar{Z}$ with the boundary contains a point F such that $h^1(F(1)) = 0$. □

Corollary 8.2. *Suppose Z is an irreducible component of $M(10)$. Then the bundle E parametrized by a general point of Z has seminatural cohomology, and Z is the closure of the irreducible open set $M(10)^{\text{sn}}$.*

Proof. The closure $\bar{Z}$ of Z meets the boundary in a nonempty subset, by Corollary 5.5. By the previous Corollary 8.1, there exists a point F in $\bar{Z}$ with $h^1(F(1)) = 0$; thus the general bundle E in Z also satisfies $h^1(E(1)) = 0$. By Proposition 6.1, the irreducible moduli space $M(10)^{\text{sn}}$ of bundles with seminatural cohomology is an open set of Z. □

Theorem 8.3. *The moduli space $M(10)$ of stable bundles of degree 1 and $c_2 = 10$, is irreducible.*

Proof. By Corollary 8.2, any irreducible component of $M(10)$ contains a dense open set parametrizing bundles with seminatural cohomology. By the main theorem of [Mestrano and Simpson 2013], there is only one such irreducible component. □

Theorem 8.4. *The full moduli space of stable torsion-free sheaves $\bar{M}(10)$ of degree 1 and $c_2 = 10$, has two irreducible components, $\overline{M(10)}$ and $M(10, 4)$ meeting along the irreducible component of usual triples in $M(10, 4)^{\text{sing}}$. These two components have the expected dimension, 20, hence the moduli space is good and connected.*

Proof. Recall that we know $M(10, 4)$ is irreducible by the results of [Mestrano and Simpson 2011]. Also $M(10)$ is irreducible. Any component has dimension ≥ 20, and by looking at the dimensions in Table 2, these are the only two possible irreducible components. Since they have dimension 20 which is the expected dimension, it follows that the moduli space is good.

It remains to be proven that these two components do indeed intersect in a nonempty subset, which then by Corollary 7.10 has to be the irreducible component of usual triples in $M(10, 4)^{\text{sing}}$. Notice that Corollary 7.10 did not say that the intersection was necessarily nonempty, since it started from the hypothesis that there was a meeting point. It is a consequence of Nijsse's connectedness theorem that the intersection is nonempty, but this may be seen more concretely as follows.

Consider the stratum $M(10,5)$. Recall from [Mestrano and Simpson 2011] that the moduli space $M(5)$ consists of bundles which fit into an exact sequence of the form

$$0 \to \mathcal{O}_X \to E \to J_{P/X}(1) \to 0,$$

such that $P = L \cap X$ for $L \subset \mathbb{P}^3$ a line. In what follows, choose L general so that P consists of 5 distinct points.

The space of extensions $\operatorname{Ext}^1(J_{P/X}(1), \mathcal{O}_X)$ is dual to $\operatorname{Ext}^1(\mathcal{O}_X, J_{P/X}(2)) = H^1(J_{P/X}(2))$. We have the exact sequence

$$H^0(\mathcal{O}_X(2)) \to H^0(\mathcal{O}_P(2)) \to H^1(J_{P/X}(2)) \to 0.$$

However, $H^0(\mathcal{O}_X(2)) = H^0(\mathcal{O}_{\mathbb{P}^3}(2))$ and the map to $H^0(\mathcal{O}_P(2))$ factors through $H^0(\mathcal{O}_L(2))$, the space of degree two forms on $L \cong \mathbb{P}^1$, which has dimension 3. Hence, the cokernel $H^1(J_{P/X}(2))$ has dimension 2. The extension classes which correspond to bundles, are the linear forms on $H^1(J_{P/X}(2))$ which don't vanish on any of the images of the lines in $H^0(\mathcal{O}_P(2))$ corresponding to the 5 different points. Since X is general, the collection of 5 points $X \cap L$ is not in a special position in $\mathbb{P}^1$, so the images of the lines are distinct in the two-dimensional space $H^1(J_{P/X}(2))$. So we can find a family of extension classes whose limiting point is an extension which vanishes on one of the lines corresponding to a point in P. This gives a degeneration towards a torsion-free sheaf with a single not locally free point, still sitting in a nontrivial extension of the above form. We conclude that the limiting bundle is still stable, so we have constructed a degeneration from a point of $M(5)$, to the single boundary stratum $M(5,4)$.

Notice that the dimension of $M(5,4)$ is bigger than that of $M(5)$, so the set of limiting points is a strict subvariety of $M(5,4)$. We have $\bar{M}(5) = M(5) \cup M(5,4)$, and we have shown that the closures of these two strata have nonempty intersection. This fact is also a consequence of the more explicit description of $\overline{M(5)}$ stated in Theorem 7.2 above (but where the proof was left to the reader).

Moving up to $c_2 = 10$, it follows that the closure of the stratum $M(10,5)$ intersects $M(10,4)$. However, $M(10,4)$ is closed, and the remaining strata of the boundary have dimension ≤ 19, so all of the other strata in the boundary, in particular $M(10,5)$, are contained in the closure of the locus of bundles $\overline{M(10)}$. Thus, $\overline{M(10,5)} \subset \overline{M(10)}$, but $M(10,4) \cap \overline{M(10,5)} \neq \varnothing$, proving that the intersection $M(10,4) \cap \overline{M(10)}$ is nonempty. □

Physics discussion. From this fact, we see that there are degenerations of stable bundles in $M(10)$, near to boundary points in $M(10,4)$. Donaldson's Yang–Mills connections then degenerate towards Uhlenbeck boundary points, connections where 6 delta-like singular instantons appear. However, these degenerations go not to all points in $M(10,4)$ but only to ones which are in the irreducible subvariety $M(10,4)^{\mathrm{sing}} \subset M(10,4)$ consisting of points on the internal spectral variety of a nonzero Higgs field $\varphi : E \to E \otimes K_X$. It gives a constraint of a global nature on the 6-tuples of singular instantons which can appear in Yang–Mills connections on a stable bundle $F \in M(10)$. It would be interesting to understand the geometry of the Higgs field which shows up, somewhat virtually, in the limit.

9. Irreducibility for $c_2 \geq 11$

Consider next the moduli space $\bar{M}(11)$ of stable torsion-free sheaves of degree one and $c_2 = 11$. The moduli space is good, of dimension 24. From Table 2, the dimensions of the boundary strata are all ≤ 23, so the set of irreducible components of $\bar{M}(11)$ is the same as the set of irreducible components of $M(11)$. Suppose Z is an irreducible component. By Corollary 5.5, Z meets the boundary in a nonempty subset of codimension 1, i.e., dimension 23. From Table 2, the only two possibilities are $M(11,10)$ and $M(11,4)$. Note that $M(11,4)$ is closed since it is the lowest stratum; it is irreducible by Li's theorem and irreducibility of $M(4)$. The stratum $M(11,10)$ is irreducible because of Theorem 8.3.

Lemma 9.1. *The intersection $M(11,4) \cap \overline{M(11,10)}$ is a nonempty subset containing, in particular, points which are torsion-free sheaves F' entering into an exact sequence of the form*

$$0 \to F' \to F \to S_x \to 0,$$

where F is a usual point of $M(10,4)^{\mathrm{sing}}$, $x \in X$ is a general point, and $F \to S_x$ is a general length one quotient.

Proof. Theorem 8.4 shows that the intersection $M(10,4) \cap \overline{M(10)}$ is nonempty. It is the unique 19-dimensional irreducible component of $M(10,4)^{\mathrm{sing}}$, containing the usual points. Starting with a general point $F \in M(10,4) \cap \overline{M(10)}$ and taking an additional general length 1 quotient S_x, the subsheaf F' gives a point in $M(11,4) \cap \overline{M(11,10)}$. □

Let $Y \subset M(10,4)$ be the unique 19-dimensional irreducible component of the singular locus $M(10,4)^{\mathrm{sing}}$. It contains a dense open set where the quotient S is a direct sum of six quotients (x_i, s_i) of length 1. Choose a quasifinite surjection $Y' \to Y$ such that (x_i, s_i) are well defined as functions $Y' \to \mathbb{P}(E)$.

Forgetting the quotients and considering only the bundle E gives a map $Y' \to M(4)$. Fix a bundle E in the image of $Y' \to M(4)$. Let Y'_E denote the fiber of Y' over E, which has dimension ≥ 17.

We claim that for any $0 \leq k \leq 5$, there exists a choice of k out of the 6 points such that the map $Y'_E \to \mathbb{P}(E)^k$ is surjective. For $k = 0$ this is automatic, so assume that

$k \leq 4$ and it is known for k; we need to show that it is true for $k+1$ points. Reorder so that the k points to be chosen, are the first ones. For a general point $q \in \mathbb{P}(E)^k$, let $Y'_{E,q}$ denote the fiber of $Y'_E \to \mathbb{P}(E)^k$ over q. We have $\dim(Y'_{E,q}) \geq 17 - 3k$. We get an injection

$$Y'_{E,q} \to \mathbb{P}(E)^{6-k}.$$

Suppose that the image mapped into a proper subvariety of each factor; then it would map into a subvariety of dimension $\leq 2(6-k)$, which would give $\dim(Y'_{E,q}) \leq 12-2k$. However, for $k \leq 4$ we have $12-2k < 17-3k$, a contradiction. Therefore, at least one of the projections must be a surjection $Y'_{E,q} \to \mathbb{P}(E)$. Adding this point to our list, gives a list of $k+1$ points such that the map $Y'_E \to \mathbb{P}(E)^{k+1}$ is surjective. This completes the induction, yielding the following lemma.

Lemma 9.2. *Suppose $Y \subset M(10, 4)$ is as above. Then for a fixed bundle $E \in M(4)$ corresponding to some points in Y, and for a general point in the fiber Y_E over E, some 5 out of the 6 quotients correspond to a general point of $\mathbb{P}(E)^5$.*

Lemma 9.3. *Suppose F is the torsion-free sheaf parametrized by a general point of Y, and let F' be defined by an exact sequence*

$$0 \to F' \to F \xrightarrow{(x_7,s_7)} S_7 \to 0,$$

where S_7 has length 1 and (x_7, s_7) is general (with respect to the choice of F) in $\mathbb{P}(E)$. Then F' has no nontrivial coobstructions: $\mathrm{Hom}(F', F'(1)) = 0$.

Proof. The space of coobstructions for the bundle E has dimension 6. Imposing a condition of compatibility with a general length-1 quotient (x_i, s_i) cuts down the dimension of the space of coobstructions by at least 1.

By Lemma 9.2 above, we may assume after reordering that the first five points $(x_1, s_1), \ldots, (x_5, s_5)$ constitute a general vector in $\mathbb{P}(E)^5$. Adding the 7-th general point given by the statement of the proposition, we obtain a general point $(x_1, s_1), \ldots, (x_5, s_5), (x_7, s_7)$ in $\mathbb{P}(E)^6$. As this 6-tuple of points is general with respect to E, it imposes vanishing on the 6-dimensional space of coobstructions, giving $\mathrm{Hom}(F', F'(1)) = 0$. $\square$

Corollary 9.4. *There exists a point*

$$F' \in \overline{M(11, 10)} \cap M(11, 4)$$

in the boundary of $\overline{M}(11)$, such that F' is a smooth point of $\overline{M}(11)$.

Proof. By Lemma 9.3, choosing a general quotient (x_7, s_7) gives a torsion-free sheaf F' with no coobstructions, hence corresponding to a smooth point of $\overline{M}(11)$. By construction we have $F' \in \overline{M(11, 10)} \cap M(11, 4)$. $\square$

Theorem 9.5. *The moduli space $\overline{M}(11)$ is irreducible.*

Proof. Suppose Z is an irreducible component. Then Z meets the boundary in a codimension 1 subset; but by looking at Table 2, there are only two possibilities: $\overline{M(11, 10)}$ and $M(11, 4)$. The coobstructions vanish for general points of $M(10, 4)$ since those correspond to 6 general quotients of length 1, and the coobstructions vanish for general points of $M(10)$ by goodness. It follows that there are no coobstructions at general points of $\overline{M(11, 10)}$ or $M(11, 4)$, so each of these is contained in at most a single irreducible component of $\bar{M}(11)$. However, in the previous corollary, there is a unique irreducible component containing F', which shows that the irreducible components containing $\overline{M(11, 10)}$ and $M(11, 4)$ must be the same. Hence, $\bar{M}(11)$ has only one irreducible component. □

Remark. Sarbeswar Pal has pointed out to us a simplified proof for $c_2 \geq 11$, avoiding the use of Lemma 9.1. He observes from the connectedness property and goodness of the moduli of torsion-free sheaves, that any change of irreducible component must occur along a codimension 1 piece of the singular locus. However, general points of the boundary components are smooth points of the full moduli space, by an easier version of the previous discussion, so we can conclude that the singular locus has codimension ≥ 2. We have nonetheless presented our original proof since it gives some additional geometrical information on the intersection of the two boundary strata.

The cases $c_2 \geq 12$ are now easy to treat.

Theorem 9.6. *For any $c_2 \geq 12$, the moduli space $\bar{M}(c_2)$ of stable torsion-free sheaves of degree* 1 *and second Chern class c_2, is irreducible.*

Proof. By Corollary 5.5, any irreducible component of $\bar{M}(c_2)$ meets the boundary in a subset of codimension 1. However, for $c_2 \geq 12$, the only stratum of codimension 1 is $M(c_2, c_2 - 1)$. By induction on c_2, starting at $c_2 = 11$, we may assume that $M(c_2, c_2 - 1)$ is irreducible. Furthermore, if E is a general point of $M(c_2 - 1)$, then E admits no coobstructions, since $M(c_2 - 1)$ is good. Hence, a general point F in $M(c_2, c_2 - 1)$, which is the kernel of a general length-1 quotient $E \to S$, doesn't admit any coobstructions either. Therefore, $\bar{M}(c_2)$ is smooth at a general point of $M(c_2, c_2 - 1)$. Thus, there is a unique irreducible component containing $M(c_2, c_2 - 1)$, which completes the proof that $\bar{M}(c_2)$ is irreducible. □

We have finished proving our main statement, Theorem 1.1 of the introduction: *for any $c_2 \geq 4$, the moduli space $M(c_2)$ of stable vector bundles of degree* 1 *and second Chern class c_2 on a very general quintic hypersurface $X \subset \mathbb{P}^3$ is nonempty and irreducible.*

For $4 \leq c_2 \leq 9$, this is shown in [Mestrano and Simpson 2011]. For $c_2 = 10$ it is Theorem 8.3, for $c_2 = 11$ it is Theorem 9.5, and $c_2 \geq 12$ it is Theorem 9.6. Note that for $c_2 \geq 16$ it is Nijsse's [1995] theorem.

It was shown in [Mestrano and Simpson 2011] that the moduli space is good for $c_2 \geq 10$ (shown by Nijsse for $c_2 \geq 13$), and from Table 1 we see that it isn't good for $4 \leq c_2 \leq 9$. The moduli space of torsion-free sheaves $\overline{M}(c_2)$ is irreducible for $c_2 \geq 11$, as may be seen by looking at the dimensions of boundary strata in Table 2. Whereas $M(4) = \overline{M}(4)$ is irreducible, the dimensions of the strata in Table 2 imply that $\overline{M}(c_2)$ has several irreducible components for $5 \leq c_2 \leq 9$, although we haven't answered the question as to their precise number. By Theorem 8.4, $\overline{M}(10)$ has two irreducible components $\overline{M(10)}$ and $M(10, 4)$.

10. An irregularity estimate

In this section we provide a correction and improvement to [Mestrano and Simpson 2011, Lemma 5.1] and hence Corollary 5.1 there. There was an error in the proof given in there.

Lemma 10.1. *Suppose X is a very general quintic hypersurface in $\mathbb{P}^3$. Suppose $s \in H^0(\mathcal{O}_X(2))$ is a section which is not the square of a section of $\mathcal{O}_X(1)$. It defines an irreducible spectral covering $Z \subset \mathrm{Tot}(K_X)$ consisting of square roots of s. Let $\tilde{Z}$ be a resolution of singularities of Z. Then the irregularity of $\tilde{Z}$ is zero, that is to say, $H^0(\tilde{Z}, \Omega^1_{\tilde{Z}}) = 0$. Hence the dimension of $\mathrm{Pic}^0(\tilde{Z})$ is zero.*

Proof. The divisor D of zeros of s is reduced since s isn't a square and in view of the fact that $\mathcal{O}_X(1)$ generates $\mathrm{Pic}(X)$. Therefore the map $Z \to X$ is ramified with simple ramification along the smooth points of D. The involution of multiplication by -1 acts in the fibers. Choose an equivariant resolution of singularities $\tilde{Z} \to Z$ with an involution $\sigma : \tilde{Z} \to \tilde{Z}$ covering the given involution of Z. The irregularity of $\tilde{Z}$ is independent of the choice of resolution, so we would like to show that $H^0(\tilde{Z}, \Omega^1_{\tilde{Z}}) = 0$.

The map $p : \tilde{Z} \to X$ induces an exact sequence

$$0 \to \mathcal{O}_X \to p_*(\mathcal{O}_{\tilde{Z}}) \to Q \to 0,$$

with Q a rank 1 torsion-free sheaf on X. The double dual Q^{**} is a line bundle L. Using the involution σ, the above exact sequence splits: Q is the anti-invariant part. Multiplying together sections of Q gives a map

$$Q \otimes Q \to \mathcal{O}_X,$$

which extends by Hartogs to a map

$$L \otimes L \to \mathcal{O}_X.$$

Look locally near a smooth point of D where X has coordinates (x, y) such that D is given by $y = 0$, and $\tilde{Z}$ has coordinates (x, z) with $y = z^2$. As a $\mathbb{C}\{x, y\}$-module, Q or equivalently L is generated by z. The image of the multiplication map is therefore the submodule generated by $z^2 = y$. It is an isomorphism outside of D, and to get an isomorphism it suffices to look off of codimension 2. This shows that

$$L \otimes L \xrightarrow{\sim} \mathcal{O}_X(-D) \cong \mathcal{O}_X(-2),$$

hence $L \cong \mathcal{O}_X(-1)$. It means that L is generated by the linear functions along the fibers of $K_X \to X$, restricted back to $\tilde{Z}$.

Consider similarly the decomposition into invariant and anti-invariant pieces

$$p_*(\Omega^1_{\tilde{Z}}) = \mathcal{F}^+ \oplus \mathcal{F}^-.$$

These sheaves are torsion-free, and we have a map $\Omega^1_X \to \mathcal{F}^+$. Again with the local coordinates x, y for X and x, z for $\tilde{Z}$ near a smooth point of D as above, we have that $\Omega^1_{\tilde{Z}}$ is generated by dx and dz. As a module over $\mathbb{C}\{x, y\}$, $\mathcal{F}^+$ is generated by dx and zdz or equivalently dx and dy. This shows that the map $\Omega^1_X \to \mathcal{F}^+$ is an isomorphism on smooth points of D. Since $\mathcal{F}^+$ is torsion-free and Ω^1_X is locally free, it follows that this map is an isomorphism. We may therefore write

$$p_*(\Omega^1_{\tilde{Z}}) = \Omega^1_X \oplus \mathcal{F}^-.$$

Consider now the map $\Omega^1_X \otimes Q \to \mathcal{F}^-$. Let $\mathcal{G} := (\mathcal{F}^-)^{**}$ be the double dual, and the previous map induces a map

$$\Omega^1_X \otimes L \to \mathcal{G}.$$

Consider again the situation at a smooth point of D using local coordinates. Note that $\mathcal{G}$ is generated by zdx and dz, whereas $\Omega^1_X \otimes L$ is generated by zdx and $zdy = z^2dz = ydz$. Recalling that $L = \mathcal{O}_X(-1)$, we get an exact sequence

$$0 \to \Omega^1_X(-1) \to \mathcal{G} \to \mathcal{B} \to 0,$$

where $\mathcal{B}$ is a sheaf supported on D, locally near the smooth points being isomorphic to $\mathcal{O}_D$. This says that $\mathcal{G}$ and $\Omega^1_X(-1)$ are related by an elementary transformation. In particular, we get

$$0 \to \mathcal{G} \to \Omega^1_X(-1)(D) = \Omega^1_X(1).$$

The irregularity of X vanishes so $H^0(\Omega^1_X) = 0$. Hence,

$$H^0(\tilde{Z}, \Omega^1_{\tilde{Z}}) \cong H^0(X, p_*\Omega^1_{\tilde{Z}}) \xrightarrow{\sim} H^0(X, \mathcal{F}^-) \hookrightarrow H^0(X, \mathcal{G}) \hookrightarrow H^0(X, \Omega^1_X(1)).$$

We have finally shown that there is an injection

$$H^0(\tilde{Z}, \Omega^1_{\tilde{Z}}) \hookrightarrow H^0(X, \Omega^1_X(1)).$$

One may show[8] the right-hand space of sections vanishes, completing the proof. □

[8] For convenience, here is the argument. The canonical exact sequence

$$0 \to \Omega^1_{\mathbb{P}^3} \to \mathcal{O}_{\mathbb{P}^3}(-1)^4 \to \mathcal{O}_{\mathbb{P}^3} \to 0$$

Therefore Corollary 5.1 of [Mestrano and Simpson 2011] holds, with the improved bound that the dimension is ≤ 9. Along the way we have answered [Mestrano and Simpson 2011, Question 5.1]: in the notation from there, $A = 0$.

11. Example on a degree 6 hypersurface

In this section we shall start in the direction of considering hypersurfaces of higher degree, and consider briefly the case of hypersurfaces of degree 6. In particular, the notation differs from that in effect previously.

Here, $X \subset \mathbb{P}^3$ is a very general hypersurface of degree 6, which will be denoted $X = X^6$ in the statements of the main corollaries, for precision. We have $K_X = \mathcal{O}_X(2)$. We consider stable rank 2 vector bundles E of degree 1 and more precisely with $\det(E) = \mathcal{O}_X(1)$, and some specified value of c_2.

Assume $h^0(E) > 0$. Then there is a section, corresponding to a morphism $s : \mathcal{O}_X \to E$. The zeros of s are in codimension 2; otherwise it would extend to $\mathcal{O}_X(1) \to E$, contradicting stability. Therefore, s fits into an exact sequence of the usual form

$$0 \to \mathcal{O}_X \to E \to J_{P/X}(1) \to 0, \tag{11-1}$$

where $P \subset X$ is a locally complete intersection subscheme of dimension 0. By the general theory, P satisfies the condition $\mathrm{CB}(L^{-1} \otimes M \otimes K_X)$, where $L = \mathcal{O}_X$ and $M = \mathcal{O}_X(1)$. In other words, P is a $\mathrm{CB}(3)$ subscheme.

Notice that $c_2(\mathcal{O}_X \oplus \mathcal{O}_X(1)) = 0$ by the product formula for Chern polynomials; therefore in the above extension, we have $c_2(E) = |P|$.

In our examples, we will consider the case $c_2 = 11$, and give two different kinds of 11-point $\mathrm{CB}(3)$ subschemes.

Before getting to these, let us note some general things about the deformation theory. Our bundle satisfies $E^* = E(-1)$, so

$$\mathrm{End}(E) = E^* \otimes E \cong E \otimes E(-1).$$

---

gives rise to

$$0 \to H^0(\Omega^1_{\mathbb{P}^3}(1)) \to H^0(\mathcal{O}^4_{\mathbb{P}^3}) \to H^0(\mathcal{O}_{\mathbb{P}^3}(1)),$$

in which the right map is an isomorphism, so $H^0(\Omega^1_{\mathbb{P}^3}(1)) = 0$. We also get $H^1(\mathbb{P}^3, \Omega^1_{\mathbb{P}^3}(-4)) = 0$; thus the exact sequence

$$0 \to \Omega^1_{\mathbb{P}^3}(-4) \to \Omega^1_{\mathbb{P}^3}(1) \to \Omega^1_{\mathbb{P}^3}(1)|_X \to 0$$

implies $H^0(\Omega^1_{\mathbb{P}^3}(1)|_X) = 0$. Now using $H^1(\mathcal{O}_X(n)) = 0$, the exact sequence

$$0 \to N^*_{X/\mathbb{P}^3}(1) = \mathcal{O}_X(-4) \to \Omega^1_{\mathbb{P}^3}(1)|_X \to \Omega^1_X(1) \to 0$$

gives $H^0(\Omega^1_X(1)) = 0$.

The decomposition $\mathrm{End}(E) = \mathrm{End}^0(E) \oplus \mathcal{O}_X$ into the trace-free plus the central part, corresponds to the decomposition

$$E \otimes E(-1) = \mathrm{Sym}^2(E)(-1) \oplus \bigwedge^2(E)(-1).$$

Let us denote for short $V := \mathrm{Sym}^2(E)(-1)$. The deformation theory of E as a bundle with fixed determinant is governed by $H^*(V)$. Notice that if E is stable, it has no endomorphisms except the scalars, so $H^0(V) = 0$. We may also apply Serre duality, noting that V is self-dual and recalling $K_X = \mathcal{O}_X(2)$. The space of infinitesimal deformations of E is

$$\mathrm{Def}(E) = H^1(V) \cong H^1(V(2))^*,$$

and the space of obstructions is

$$\mathrm{Obs}(E) = H^2(V) \cong H^0(V(2))^*.$$

Let $2P$ denote the subscheme defined by the square of the ideal of P, so $J_{2P/X} = (J_{P/X})^2$. We have an exact sequence

$$0 \to E(-1) \to V \to J_{2P/X}(1) \to 0, \tag{11-2}$$

and hence

$$0 \to E(1) \to V(2) \to J_{2P/X}(3) \to 0. \tag{11-3}$$

Points on the rational normal cubic. The first case is when $C \subset \mathbb{P}^3$ is a general rational normal cubic, and $P \subset X \cap C$ is a collection of 11 points. This exists since $C \cap X$ consists of 18 distinct points and we may choose 11 of them.

Notice that $C \cong \mathbb{P}^1$ and $\mathcal{O}_{\mathbb{P}^3}(1)|_C = \mathcal{O}_C(3p)$ for any point $p \in C$, that is to say it is a line bundle of degree 3. Thus, $\mathcal{O}_{\mathbb{P}^3}(3)|_C = \mathcal{O}_C(9p)$ has degree 9. If $P' \subset P$ is any collection of 10 points, a section of $\mathcal{O}_{\mathbb{P}^3}(3)$ vanishing on P' must vanish on C, hence it must vanish on P. The sections of $\mathcal{O}_X(3)$ are all restrictions of sections of $\mathcal{O}_{\mathbb{P}^3}(3)$, so this proves that P satisfies the property $\mathrm{CB}(3)$.

The space of extensions of $J_{P/X}(1)$ by $\mathcal{O}_X$ is dual to $H^1(J_{P/X}(3))$, which in turn is the cokernel of

$$H^0(\mathcal{O}_X(3)) \to H_0(P, \mathcal{O}_P(3)) \cong \mathbb{C}^{11}. \tag{11-4}$$

As we have seen above, a section of $H^0(\mathcal{O}_X(3))$ vanishing on P corresponds to a section of $H^0(\mathcal{O}_{\mathbb{P}^3}(3))$ vanishing on C. One may calculate by hand that the map

$$\mathbb{C}^{20} = H^0(\mathcal{O}_{\mathbb{P}^3}(3)) \to H^0(\mathcal{O}_C(9p)) = \mathbb{C}^{10}$$

is surjective. Indeed, the image of $H^0(\mathcal{O}_{\mathbb{P}^3}(1))$ consists of the sections which may be written as $1, t, t^2, t^3$ for an affine coordinate t on $C \cong \mathbb{P}^1$ with pole at the point p. Then, monomials of degree 3 in these sections give all of the monomials $1, t, \dots, t^9$.

From this surjectivity we get that the kernel is $\mathbb{C}^{10}$. Thus, the kernel of the map (11-4) is $\mathbb{C}^{10}$ so the image of the map also has dimension 10. Finally, we get that the cokernel of (11-4) has dimension 1. We have shown that $\mathrm{Ext}^1(J_{P/X}, \mathcal{O}_X)$ has dimension 1. Therefore, a given subscheme P gives rise to only one bundle since scaling of the extension class doesn't change the isomorphism class of the bundle.

For the other direction, we claim that $h^0(E) = 1$. Consider the exact sequence

$$0 \to H^0(\mathcal{O}_X) \to H^0(E) \to H^0(J_{P/X}(1)) \to H^1(\mathcal{O}_X) = 0.$$

Given a section of $H^0(\mathcal{O}_X(1))$ vanishing on P, it comes from a section of $H^0(\mathcal{O}_{\mathbb{P}^3}(1))$ which, by the same argument as previously, vanishes on C. If the section is nonzero, that would say that C is contained in a plane, which however is not the case. Therefore, $H^0(J_{P/X}(1)) = 0$ and $\mathbb{C} \cong H^0(\mathcal{O}_X) \xrightarrow{\sim} H^0(E)$. We get $h^0(E) = 1$ as claimed.

In particular, for a given bundle E, the choice of section s is unique up to a scalar, so the subscheme P is uniquely determined.

By these arguments, we conclude that the space of bundles E in this case is isomorphic to the space of choices of subscheme $P \subset C \cap X$.

Now, given $P \subset C \cap X$ of length 11, we claim that C is the only rational normal curve passing through P. Indeed, suppose C' were another one. Note that C' is cut out by conics. If $Q \subset \mathbb{P}^3$ is a conic containing C' then $Q \cap C$ is either equal to C, or has length 6; the latter case can't happen so $C \subset Q$. Thus, any conic containing C' also contains C, which shows that $C = C'$.

The dimension of the space of subschemes P in this case is therefore equal to the dimension of the space $\mathrm{PGL}(4)/\mathrm{PGL}(2)$ of rational normal cubic curves, which is $15 - 3 = 12$. This completes the proof of the following proposition:

Proposition 11.1. *The space of bundles E fitting into an exact sequence of the form (11-1), where P is a length 11 subscheme of $C \cap X$ for C a rational normal cubic in $\mathbb{P}^3$, has dimension 12.*

Lemma 11.2. *Suppose E is a bundle fitting into an exact sequence of the form (11-1), where P is a length 11 subscheme of $C \cap X$ for C a general rational normal cubic in $\mathbb{P}^3$. Then $h^1(\mathrm{End}^0(E)) = h^1(V) = 12$.*

Proof. Use the exact sequence (11-2). The first step is to calculate $h^1(E(-1))$. Note that (11-1) gives the following sequence, using that $h^1(\mathcal{O}_X(n)) = 0$ for any n as well as $H^2(J_{P/X}(n)) = H^2(\mathcal{O}_X(n))$:

$$0 \to H^1(E(-1)) \to H^1(J_{P/X}) \to H^2(\mathcal{O}_X(-1)) \to H^2(E(-1)) \to H^2(\mathcal{O}_X) \to 0.$$

Now $H^2(E(-1))$ is dual to $H^0(E(2))$ which itself fits into the sequence

$$0 \to H^0(\mathcal{O}_X(2)) \to H^0(E(2)) \to H^0(J_{P/X}(3)) \to 0.$$

We have $H^0(J_{P/X}(3)) \cong H^0(J_{C/\mathbb{P}^3}(3)) = \ker(H^0(\mathcal{O}_{\mathbb{P}^3}(3)) \to H^0(\mathcal{O}_C(9p)))$. The latter map is surjective from $\mathbb{C}^{20}$ to $\mathbb{C}^{10}$ so its kernel has dimension 10. This gives $h^0(J_{P/X}(3)) = 10$. Also $h^0(\mathcal{O}_X(2)) = 10$ so $h^2(E(-1)) = h^0(E(2)) = 20$. We have $h^2(\mathcal{O}_X) = h^0(\mathcal{O}_X(2)) = 10$ and $h^2(\mathcal{O}_X(-1)) = h^0(\mathcal{O}_X(3)) = 20$. Finally, $H^1(J_{P/X})$ is just $\mathbb{C}^{11}$ modulo $H^0(\mathcal{O}_X) = \mathbb{C}$ so $h^1(J_{P/X}) = 10$. The alternating sum from the above sequence vanishes, saying now that

$$h^1(E(-1)) - 10 + 20 - 20 + 10 = 0,$$

so $h^1(E(-1)) = 0$.

The long exact sequence associated to (11-2) starting with $H^1(E(-1)) = 0$ now gives

$$0 \to H^1(V) \to H^1(J_{2P/X}(1)) \to H^2(E(-1)) \to H^2(V) \to H^2(\mathcal{O}_X(1)) \to 0.$$

As we have seen above, $h^2(E(-1)) = 20$. It is also easy to see that $h^0(J_{2P/X}(1)) = 0$ (we will in fact see this for $J_{2P/X}(3)$ below), so noting that the length of $2P$ is 33 we get $h^1(J_{2P/X}(1)) = 33 - h^0(\mathcal{O}_X(1)) = 29$. Putting these together and using $h^2(\mathcal{O}_X(1)) = h^0(\mathcal{O}_X(1)) = 4$ we get

$$h^1(V) - 29 + 20 - h^2(V) + 4 = 0,$$

so $h^1(V) - h^2(V) = 5$. This is the expected dimension of the moduli space.

Next, by duality $h^2(V) = h^0(V(2))$ which we can calculate using the sequence (11-3). We have

$$0 \to H^0(E(1)) \to H^0(V(2)) \to H^0(J_{2P/X}(3)).$$

We claim that $H^0(J_{2P/X}(3)) = 0$. To see this, consider a smooth quadric surface $Q \subset \mathbb{P}^3$ containing C. We have $Q \cong \mathbb{P}^1 \times \mathbb{P}^1$ and C is a divisor of bidegree $(1, 2)$ on Q. On the other hand, $\mathcal{O}_Q(1)$ has bidegree $(1, 1)$. Suppose we have a section u of $H^0(\mathcal{O}_X(3)) = H^0(\mathcal{O}_{\mathbb{P}^3}(3))$ vanishing on the $2P$ (recall that $2P$ is the subscheme of X defined by the square of the ideal of P). We have seen already above that vanishing on P implies that it vanishes on C. Therefore $u|_Q$ is a section of the bundle of bidegree $(3, 3) - (1, 2) = (2, 1)$. The intersection of $2P$ with Q consists of a collection of double points transverse to C at the points of P, so it imposes again a single condition on the section u considered as a section of $\mathcal{O}_Q(2, 1)$. The restriction of $\mathcal{O}_Q(2, 1)$ to C is a line bundle on $C \cong \mathbb{P}^1$ of degree equal to the intersection number $(2, 1).(1, 2) = 5$. Therefore, a section of $\mathcal{O}_Q(2, 1)$ which vanishes on 11 points has to vanish. This says that our section of bidegree $(2, 1)$ again vanishes on C, so it is a section of a bundle of bidegree $(1, -1)$; but that is not effective so this section has to vanish. This proves that our section $u|_Q$ vanishes. Therefore, u may be viewed as a section of $\mathcal{O}_{\mathbb{P}^3}(3)(-Q) = \mathcal{O}_{\mathbb{P}^3}(1)$. The remaining pieces of the double points composing $2P$ give conditions of vanishing again at all the points of P for this section of $\mathcal{O}_{\mathbb{P}^3}(1)$, but as C is not contained in a plane, it implies that the section vanishes. This completes the proof that $H^0(J_{2P/X}(3)) = 0$. We conclude from the previous exact sequence that

$$h^2(V) = h^0(V(2)) = h^0(E(1)).$$

Now use the sequence

$$0 \to H^0(\mathcal{O}_X(1)) \to H^0(E(1)) \to H^0(J_{P/X}(2)) \to 0.$$

As usual, $H^0(J_{P/X}(2))$ is isomorphic to the kernel of the restriction map

$$\mathbb{C}^{10} = H^0(\mathcal{O}_X(2)) \to H^0(\mathcal{O}_C(6p)) = \mathbb{C}^7$$

and this restriction map is surjective, so its kernel has dimension 3. We get

$$h^0(E(1)) = 4 + 3 = 7.$$

Thus, $h^2(V) = 7$, and putting this together with the formula that the expected dimension is 5, we have finally shown $h^1(V) = 12$. This proves the lemma. $\square$

Even though there is a 7-dimensional obstruction space, we have constructed a 12-dimensional family; it follows that all of the obstructions vanish and a general point lies in a generically smooth irreducible component of dimension 12.

Corollary 11.3. *The space of bundles E fitting into an exact sequence of the form (11-1), where P is a length 11 subscheme of $C \cap X$ for C a rational normal cubic in $\mathbb{P}^3$, consists of a single 12-dimensional generically smooth irreducible component of the moduli space $M_{X^6}(2, 1, 11)$ of stable bundles of rank 2, degree 1 and $c_2 = 11$ on our degree 6 hypersurface $X = X^6$.*

Proof. In order to understand how many irreducible components are produced by this construction, we should investigate the monodromy of the set of choices of 11 out of the 18 points of $C \cap X$, as C moves. A choice of 6 points determines the rational normal cubic C, so any 6 points can be moved to any 6 other ones. Therefore, the monodromy action is 6-tuply transitive. On the other hand, it contains a transposition, since we can move C around a choice of curve that is simply tangent to X at one point. Therefore, the monodromy group contains all transpositions, hence it is the full symmetric group on 18 elements. It acts transitively on the set of choices of 11 out of the 18 intersection points, so our construction produces a single irreducible component. $\square$

Points on a plane. The other construction we have found for CB(3) subschemes is to take 11 points in a plane. Let H be a plane in general position with respect to X, and let $Y = X \cap H$. Let P consist of a general collection of 11 points in Y.

Suppose $P' \subset P$ is a subset of 10 points. The map $H^0(\mathcal{O}_H(3)) \to H^0(\mathcal{O}_Y(3))$ is injective (since Y is a curve of degree 6 in the plane H), so a general collection of 10 points imposes independent conditions on $H^0(\mathcal{O}_H(3))$. As $h^0(\mathcal{O}_H(3)) = 10$, it means that $H^0(J_{P'/H}(3)) = 0$, hence a section of $H^0(\mathcal{O}_{\mathbb{P}^3}(3))$ vanishing on P', has to vanish on H. In particular it vanishes on P, proving the CB(3) property for P. This also gives the formula

$$H^0(J_{P/X}(3)) \cong H^0(\mathcal{O}_X(2)) = \mathbb{C}^{10}.$$

Consider next the space of choices of extension (11-1). As

$$\dim(\operatorname{Ext}^1(J_{P/X}(1), \mathcal{O}_X)) = h^1(J_{P/X}(3)) = 11 - 20 + h^0(J_{P/X}(3)) = 1,$$

whereas scalar multiples of an extension class give the same bundle, it means that for a given P there is a single corresponding bundle. On the other hand, we have $h^0(J_{P/X}(1)) = 1$ since P is contained in a plane, so $h^0(E) = 2$. This means that for a given bundle E, the space of choices of section s (modulo scaling) leading to the subscheme P, has dimension 1. Hence the dimension of the space of bundles obtained by this construction is one less than the dimension of the space of subschemes:

$$\dim\{E\} = \dim\{P\} - 1.$$

Count now the dimension of the space of choices of P: there is a three-dimensional space of choices of the plane H, and for each one we have an 11-dimensional space of choices of the subscheme P of 11 points in Y. This gives $\dim\{P\} = 3 + 11 = 14$, so $\dim\{E\} = 13$. Altogether, we have constructed a 13-dimensional family of stable bundles. It follows that this family must be in at least one irreducible component distinct from the 12-dimensional component constructed above. This proves the following theorem:

Theorem 11.4. *For a very general degree* 6 *hypersurface* $X^6 \subset \mathbb{P}^3$, *the moduli space* $M_{X^6}(2, 1, 11)$ *contains a generically smooth* 12*-dimensional component from Corollary 11.3, and contains at least one irreducible component of dimension* ≥ 13. *In particular, it is not irreducible.*

The general bundle in our 13-dimensional family may be viewed as an elementary transformation [Maruyama 1973; 1982]. A general line bundle L of degree 11 on Y has a 2-dimensional space of sections and the two sections generate L. If $j : Y \hookrightarrow X$ denotes the inclusion then we get a bundle E, elementary transformation of $\mathcal{O}_X^2$, fitting into exact sequences

$$\begin{aligned} 0 \to E(-1) \to \mathcal{O}_X^2 \to &\quad j_*(L) \quad \to 0, \\ 0 \to \quad \mathcal{O}_X^2 \quad \to \quad E \quad \to &\; j_*(L^*)(1) \to 0. \end{aligned}$$

This shows that E determines Y and L. Since Y has genus 10, the space of choices of hyperplane plus choice of L has dimension $3+10=13$. One may see that these bundles are the same as the previous ones, indeed the zeros of a section of our elementary transformation E are the same as those of the corresponding section of L. This gives an alternate canonical viewpoint on our second construction of bundles that should be useful for understanding the obstruction map.

We conjecture that the rational normal case and the planar case cover all of $M_{X^6}(2,1,11)$. More precisely:

Conjecture 11.5. The 13-dimensional family constructed in the present subsection constitutes a full irreducible component of $M_{X^6}(2,1,11)$; this component is nonreduced and obstructed. Together with the 12-dimensional generically smooth component constructed in the previous subsection, these are the only irreducible components of $M_{X^6}(2,1,11)$. In particular, $h^0(E)>0$ for any stable bundle with $c_2=11$.

There doesn't seem to be an easy direct proof of the property $h^0(E)>0$. The Euler characteristic consideration does give $h^0(E(1))>0$ so any E has to be in an extension of $\mathcal{O}(-1)$ by $J_{P/X}(2)$ with P satisfying CB(5). If this conjecture is true, it would imply that any CB(5) subscheme of length 21 contained in X^6, would have to be contained in a quadric hypersurface. We didn't find a proof of that, but we couldn't find any length-21 subschemes of X^6 satisfying CB(5) that weren't contained in quadric hypersurfaces either, leading to the conjecture.

References

[Arrondo 2007] E. Arrondo, "A home-made Hartshorne–Serre correspondence", *Rev. Mat. Complut.* **20**:2 (2007), 423–443. MR Zbl

[Barth 1977] W. Barth, "Moduli of vector bundles on the projective plane", *Invent. Math.* **42** (1977), 63–91. MR Zbl

[Briançon et al. 1981] J. Briançon, M. Granger, and J.-P. Speder, "Sur le schéma de Hilbert d'une courbe plane", *Ann. Sci. École Norm. Sup.* (4) **14**:1 (1981), 1–25. MR Zbl

[Eisenbud et al. 1996] D. Eisenbud, M. Green, and J. Harris, "Cayley–Bacharach theorems and conjectures", *Bull. Amer. Math. Soc.* (*N.S.*) **33**:3 (1996), 295–324. MR Zbl

[Ellingsrud and Lehn 1999] G. Ellingsrud and M. Lehn, "Irreducibility of the punctual quotient scheme of a surface", *Ark. Mat.* **37**:2 (1999), 245–254. MR Zbl

[Gieseker 1977] D. Gieseker, "On the moduli of vector bundles on an algebraic surface", *Ann. of Math.* (2) **106**:1 (1977), 45–60. MR Zbl

[Gieseker 1988] D. Gieseker, "A construction of stable bundles on an algebraic surface", *J. Differential Geom.* **27**:1 (1988), 137–154. MR Zbl

[Gieseker and Li 1994] D. Gieseker and J. Li, "Irreducibility of moduli of rank-2 vector bundles on algebraic surfaces", *J. Differential Geom.* **40**:1 (1994), 23–104. MR Zbl

[Gómez 1997] T. L. Gómez de Quiroga, *Irreducibility of the moduli space of vector bundles on surfaces and Brill–Noether theory on singular curves*, Ph.D. thesis, Princeton University, 1997, https://search.proquest.com/docview/304411614. Revised version at arXiv alg-geom/9710029.

[Griffiths and Harris 1978] P. Griffiths and J. Harris, "Residues and zero-cycles on algebraic varieties", *Ann. of Math.* (2) **108**:3 (1978), 461–505. MR Zbl

[Hartshorne 1962] R. Hartshorne, "Complete intersections and connectedness", *Amer. J. Math.* **84** (1962), 497–508. MR Zbl

[Huybrechts and Lehn 1997] D. Huybrechts and M. Lehn, *The geometry of moduli spaces of sheaves*, Aspects of Mathematics **E31**, Friedr. Vieweg & Sohn, Braunschweig, 1997. MR Zbl

[Langer 2008] A. Langer, "Lectures on torsion-free sheaves and their moduli", pp. 69–103 in *Algebraic cycles, sheaves, shtukas, and moduli*, edited by P. Pragacz, Birkhäuser, Basel, 2008. MR Zbl

[Li 1993] J. Li, "Algebraic geometric interpretation of Donaldson's polynomial invariants", *J. Differential Geom.* **37**:2 (1993), 417–466. MR Zbl

[Maruyama 1973] M. Maruyama, "On a family of algebraic vector bundles", pp. 95–146 in *Number theory, algebraic geometry and commutative algebra*, edited by Y. Kusunoki et al., Kinokuniya, Tokyo, 1973. MR Zbl

[Maruyama 1975] M. Maruyama, "Stable vector bundles on an algebraic surface", *Nagoya Math. J.* **58** (1975), 25–68. MR Zbl

[Maruyama 1982] M. Maruyama, "Elementary transformations in the theory of algebraic vector bundles", pp. 241–266 in *Algebraic geometry* (La Rábida, 1981), edited by J. M. Aroca et al., Lecture Notes in Math. **961**, Springer, 1982. MR Zbl

[Mestrano 1997] N. Mestrano, "Sur les espaces de modules des fibrés vectoriels de rang deux sur des hypersurfaces de $\mathbb{P}^3$", *J. Reine Angew. Math.* **490** (1997), 65–79. MR Zbl

[Mestrano and Simpson 2011] N. Mestrano and C. Simpson, "Obstructed bundles of rank two on a quintic surface", *Internat. J. Math.* **22**:6 (2011), 789–836. MR Zbl

[Mestrano and Simpson 2013] N. Mestrano and C. Simpson, "Seminatural bundles of rank two, degree one, and $c_2 = 10$ on a quintic surface", *Kyoto J. Math.* **53**:1 (2013), 155–195. MR Zbl

[Mestrano and Simpson 2016] N. Mestrano and C. Simpson, "Moduli of sheaves", pp. 77–172 in *Development of moduli theory* (Kyoto, 2013), edited by O. Fujino et al., Adv. Stud. Pure Math. **69**, Math. Soc. Japan, Tokyo, 2016. MR Zbl

[Mukai 1984] S. Mukai, "Symplectic structure of the moduli space of sheaves on an abelian or $K3$ surface", *Invent. Math.* **77**:1 (1984), 101–116. MR Zbl

[Nijsse 1995] P. G. J. Nijsse, "The irreducibility of the moduli space of stable vector bundles of rank 2 on a quintic in $\mathbb{P}^3$", preprint W95-02, Leiden University, 1995. arXiv

[O'Grady 1993] K. G. O'Grady, "The irreducible components of moduli spaces of vector bundles on surfaces", *Invent. Math.* **112**:3 (1993), 585–613. MR Zbl

[O'Grady 1996] K. G. O'Grady, "Moduli of vector bundles on projective surfaces: some basic results", *Invent. Math.* **123**:1 (1996), 141–207. MR Zbl

[Reider 1988] I. Reider, "Vector bundles of rank 2 and linear systems on algebraic surfaces", *Ann. of Math.* (2) **127**:2 (1988), 309–316. MR Zbl

[Sawant 2011] A. P. Sawant, "Hartshorne's connectedness theorem", preprint, Tata Institute of Fundamental Research, 2011, http://tinyurl.com/sawantarchive.

[Yoshioka 1997] K. Yoshioka, "An application of exceptional bundles to the moduli of stable sheaves on a $K3$ surface", preprint, 1997. arXiv

[Yoshioka 1999] K. Yoshioka, "Irreducibility of moduli spaces of vector bundles on $K3$ surfaces", preprint, 1999. arXiv

[Yoshioka 2001] K. Yoshioka, "Moduli spaces of stable sheaves on abelian surfaces", *Math. Ann.* **321**:4 (2001), 817–884. MR Zbl

[Zuo 1991] K. Zuo, "Generic smoothness of the moduli spaces of rank two stable vector bundles over algebraic surfaces", *Math. Z.* **207**:4 (1991), 629–643. MR Zbl

Received January 25, 2017. Revised September 7, 2017.

NICOLE MESTRANO
LABORATOIRE J. A. DIEUDONNÉ
UNIVERSITÉ CÔTE D'AZUR
NICE
FRANCE
nicole@math.unice.fr

CARLOS SIMPSON
LABORATOIRE J. A. DIEUDONNÉ
UNIVERSITÉ CÔTE D'AZUR
NICE
FRANCE
carlos@unice.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.173

A CAPILLARY SURFACE WITH NO RADIAL LIMITS

COLM PATRIC MITCHELL

In 1996, Kirk Lancaster and David Siegel investigated the existence and behavior of radial limits at a corner of the boundary of the domain of solutions of capillary and other prescribed mean curvature problems with contact angle boundary data. They provided an example of a capillary surface in a unit disk D which has no radial limits at $(0, 0) \in \partial D$. In their example, the contact angle, γ, cannot be bounded away from zero and π. Here we consider a domain Ω with a convex corner at $(0, 0)$ and find a capillary surface $z = f(x, y)$ in $\Omega \times \mathbb{R}$ which has no radial limits at $(0, 0) \in \partial\Omega$ such that γ is bounded away from 0 and π.

Let Ω be a domain in $\mathbb{R}^2$ with locally Lipschitz boundary and $\mathcal{O} = (0, 0) \in \partial\Omega$ such that $\partial\Omega \setminus \{\mathcal{O}\}$ is a C^4 curve and $\Omega \subset B_1(0, 1)$, where $B_\delta(\mathcal{N})$ is the open ball in $\mathbb{R}^2$ of radius δ about $\mathcal{N} \in \mathbb{R}^2$. Denote the unit exterior normal to Ω at $(x, y) \in \partial\Omega$ by $\nu(x, y)$ and let polar coordinates relative to $\mathcal{O}$ be denoted by r and θ. We shall assume there exist $\delta^* \in (0, 2)$ and $\alpha \in (0, \frac{\pi}{2})$ such that $\partial\Omega \cap B_{\delta^*}(\mathcal{O})$ consists of the line segments

$$\partial^+\Omega^* = \{(r\cos(\alpha), r\sin(\alpha)) : 0 \le r \le \delta^*\}$$

and

$$\partial^-\Omega^* = \{(r\cos(-\alpha), r\sin(-\alpha)) : 0 \le r \le \delta^*\}.$$

Set $\Omega^* = \Omega \cap B_{\delta^*}(\mathcal{O})$. Let $\gamma : \partial\Omega \to [0, \pi]$ be given. Let $(x^\pm(s), y^\pm(s))$ be arclength parametrizations of $\partial^\pm\Omega$ with $(x^+(0), y^+(0)) = (x^-(0), y^-(0)) = (0, 0)$ and set $\gamma^\pm(s) = \gamma(x^\pm(s), y^\pm(s))$.

Consider the capillary problem of finding a function $f \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ satisfying

$$\operatorname{div}(Tf) = \tfrac{1}{2} f \quad \text{in } \Omega \tag{1}$$

$$Tf \cdot \nu = \cos(\gamma) \quad \text{on } \partial\Omega \setminus \{\mathcal{O}\}, \tag{2}$$

where $Tf = \nabla f / \sqrt{1 + |\nabla f|^2}$. We are interested in the existence of the radial limits $Rf(\cdot)$ of a solution f of (1) and (2), where

$$Rf(\theta) = \lim_{r \to 0^+} f(r\cos\theta, r\sin\theta), \qquad -\alpha < \theta < \alpha$$

MSC2010: 35B40, 35J93, 53A10.
Keywords: capillary surfaces, PDE, Concus–Finn conjecture.

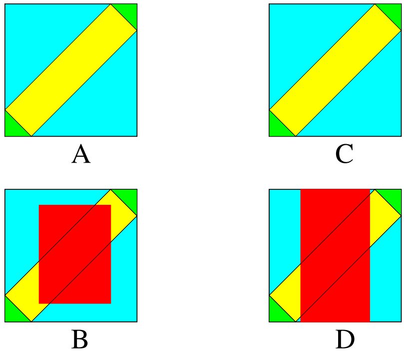


Figure 1. The Concus–Finn rectangle (A and C) with regions R (yellow), $D_2^{\pm}$ (blue) and $D_1^{\pm}$ (green); the restrictions on γ in [Lancaster and Siegel 1996] (red region in B) and in [Crenshaw et al. 2017] (red region in D).

and $Rf(\pm\alpha) = \lim_{\partial^{\pm}\Omega^{*} \ni \boldsymbol{x} \to \mathcal{O}} f(\boldsymbol{x})$, $\boldsymbol{x} = (x, y)$, which are the limits of the boundary values of f on the two sides of the corner if these exist.

Proposition 1 [Crenshaw et al. 2017]. *Let f be a bounded solution to (1) satisfying (2) on $\partial^{\pm}\Omega^{*} \setminus \{\mathcal{O}\}$ which is discontinuous at $\mathcal{O}$. If $\alpha > \pi/2$ then $Rf(\theta)$ exists for all $\theta \in (-\alpha, \alpha)$. If $\alpha \le \pi/2$ and there exist constants $\underline{\gamma}^{\pm}, \bar{\gamma}^{\pm}$, $0 \le \underline{\gamma}^{\pm} \le \bar{\gamma}^{\pm} \le \pi$, satisfying*

$$\pi - 2\alpha < \underline{\gamma}^{+} + \underline{\gamma}^{-} \le \bar{\gamma}^{+} + \bar{\gamma}^{-} < \pi + 2\alpha$$

so that $\underline{\gamma}^{\pm} \le \gamma^{\pm}(s) \le \bar{\gamma}^{\pm}$ for all s, $0 < s < s_0$, for some s_0, then again $Rf(\theta)$ exists for all $\theta \in (-\alpha, \alpha)$.

Lancaster and Siegel [1996] proved this theorem with the additional restriction that γ be bounded away from 0 and π; Figure 1 illustrates these cases.

They also proved the following:

Proposition 2 [Lancaster and Siegel 1996, Theorem 3]. *Let Ω be the disk of radius 1 centered at $(1, 0)$. Then there exists a solution to $Nf = \frac{1}{2}f$ in Ω, $|f| \le 2$, $f \in C^2(\Omega) \cap C^1(\bar{\Omega} \setminus O)$, $O = (0, 0)$ such that no radial limits $Rf(\theta)$ exist ($\theta \in [-\pi/2, \pi/2]$).*

In this case, $\alpha = \frac{\pi}{2}$; if γ is bounded away from 0 and π, then Proposition 1 would imply that $Rf(\theta)$ exists for each $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and therefore the contact angle $\gamma = \cos^{-1}(Tf \cdot \nu)$ in Proposition 2 is not bounded away from 0 and π.

In our case, the domain Ω has a convex corner of size 2α at $\mathcal{O}$ and we wish to investigate the question of whether an example like that in Proposition 2 exists in this case when γ is bounded away from 0 and π. In terms of the Concus–Finn rectangle, the question is whether, given $\epsilon > 0$, there is a solution $f \in C^2(\Omega) \cap C^1(\bar{\Omega} \setminus \{\mathcal{O}\})$ of

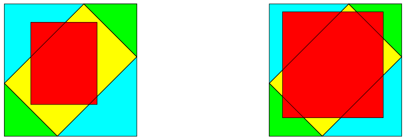


Figure 2. The Concus–Finn rectangle. When γ remains in the red region in E, $Rf(\cdot)$ exists; γ in Theorem 1 remains in the red region in F.

(1) and (2) such that no radial limits $Rf(\theta)$ exist ($\theta \in [-\alpha, \alpha]$) and $\left|\gamma - \frac{\pi}{2}\right| \le \alpha + \epsilon$; this is illustrated in Figure 2.

Theorem 1. *For each $\epsilon > 0$, there is a domain Ω as described above and a solution $f \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ of (1) such that the contact angle*

$$\gamma = \cos^{-1}(Tf \cdot \nu) : \partial\Omega \setminus \{\mathcal{O}\} \to [0, \pi]$$

satisfies $\left|\gamma - \frac{\pi}{2}\right| \le \alpha + \epsilon$ and there exists a sequence $\{r_j\}$ in $(0, 1)$ with $\lim_{j\to\infty} r_j = 0$ such that

$$(-1)^j f(r_j, 0) > 1 \quad \textit{for each } j \in \mathbb{N}.$$

Assuming Ω and γ are symmetric with respect to the line $\{(x, 0) : x \in \mathbb{R}\}$, this implies that no radial limit

$$Rf(\theta) \overset{\text{def}}{=} \lim_{r\downarrow 0} f(r\cos(\theta), r\sin(\theta)) \tag{3}$$

exists for any $\theta \in [-\alpha, \alpha]$.

We remark that our theorem is an extension of [Lancaster and Siegel 1996, Theorem 3] to contact angle data in a domain with a convex corner. As in [Lancaster 1989; Lancaster and Siegel 1996], we first state and prove a localization lemma; this is analogous to [Lancaster 1989, Lemma] and [Lancaster and Siegel 1996, Lemma 2].

Lemma 1. *Let $\Omega \subseteq \mathbb{R}^2$ be as above, $\epsilon > 0$, $\eta > 0$ and $\gamma_0 : \partial\Omega \setminus \{\mathcal{O}\} \to [0, \pi]$ such that $\left|\gamma_0 - \frac{\pi}{2}\right| \le \alpha + \epsilon$. For each $\delta \in (0, 1)$ and $h \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ which satisfies (1) and (2) with $\gamma = \gamma_0$, there exists a solution*

$$g \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$$

of (1) such that $\lim_{\overline{\Omega}\ni(x,y)\to(0,0)} g(x, y) = +\infty$,

$$\sup_{\Omega_\delta} |g - h| < \eta \quad \textit{and} \quad \left|\gamma_g - \frac{\pi}{2}\right| \le \alpha + \epsilon, \tag{4}$$

where $\Omega_\delta = \overline{\Omega} \setminus B_\delta(\mathcal{O})$ and $\gamma_g = \cos^{-1}(Tg \cdot \nu) : \partial\Omega \setminus \{\mathcal{O}\} \to [0, \pi]$ is the contact angle which the graph of g makes with $\partial\Omega \times \mathbb{R}$.

Proof. Let $\epsilon, \eta, \delta, \Omega, h$ and γ_0 be given. For $\beta \in (0,\delta)$, let $g_\beta \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ satisfy (1) and (2) with $\gamma = \gamma_\beta$, where

$$\gamma_\beta = \begin{cases} \frac{\pi}{2} - \alpha - \epsilon & \text{on } \overline{B_\beta(\mathcal{O})} \\ \gamma_0 & \text{on } \overline{\Omega} \setminus B_\beta(\mathcal{O}). \end{cases}$$

As in the proof of [Lancaster and Siegel 1996, Theorem 3], g_β converges to h pointwise and uniformly in the C^1 norm on $\overline{\Omega}_\delta$ as β tends to zero. Fix $\beta > 0$ small enough that $\sup_{\Omega_\delta} |g - h| < \eta$.

Set $\Sigma = \{(r\cos(\theta), r\sin(\theta)) : r > 0, -\alpha \le \theta \le \alpha\}$. Now define $w : \Sigma \to \mathbb{R}$ by

$$w(r\cos\theta, r\sin\theta) = \frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{k\kappa r},$$

where $k = \sin\alpha \sec\left(\frac{\pi}{2} - \alpha - \epsilon\right) = \sin\alpha \csc(\alpha + \epsilon)$. As in [Concus and Finn 1970], there exists a $\delta_1 > 0$ such that $\operatorname{div}(Tw) - \frac{1}{2}w \ge 0$ on $\Sigma \cap B_{\delta_1}(\mathcal{O})$, $Tw \cdot \nu = \cos(\frac{\pi}{2} - \alpha - \epsilon)$ on $\partial\Sigma \cap B_{\delta_1}(\mathcal{O})$, and $\lim_{r \to 0^+} w(r\cos\theta, r\sin\theta) = \infty$ for each $\theta \in [-\alpha, \alpha]$. We may assume $\delta_1 \le \delta^*$. Let

$$M = \sup_{\Omega \cap \partial B_{\delta_1}(\mathcal{O})} |w - g_\beta| \quad \text{and} \quad w_\beta = w - M.$$

Since $\operatorname{div}(Tw_\beta) - \frac{1}{2}w_\beta \ge \frac{M}{2} \ge 0 = \operatorname{div}(Tg_\beta) - \frac{1}{2}g_\beta$ in $\Omega \cap B_{\delta_1}(\mathcal{O})$, $w_\beta \le g_\beta$ on $\Omega \cap \partial B_{\delta_1}(\mathcal{O})$ and $Tg_\beta \cdot \nu \ge Tw_\beta \cdot \nu$ on $\partial\Omega \cap B_{\delta_1}(\mathcal{O})$, we see that $g_\beta \ge w_\beta$ on $\Omega \cap \partial B_{\delta_1}(\mathcal{O})$. □

We may now prove Theorem 1.

Proof. We shall construct a sequence f_n of solutions of (1) and a sequence $\{r_n\}$ of positive real numbers such that $\lim_{n\to\infty} r_n = 0$, $f_n(x, y)$ is even in y and

$$(-1)^j f_n(r_j, 0) > 1 \qquad \text{for each } j = 1, \dots, n.$$

Let $\gamma_0 = \frac{\pi}{2}$ and $f_0 = 0$. Set $\eta_1 = 1$ and $\delta_1 = \delta_0$. From Lemma 1, there exists $f_1 \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ which satisfies (1) such that $\sup_{\Omega_{\delta_1}} |f_1 - f_0| < \eta_1$, $\left|\gamma_1 - \frac{\pi}{2}\right| \le \alpha + \epsilon$ and $\lim_{\Omega \ni (x,y) \to \mathcal{O}} f_1(x, y) = -\infty$, where $\gamma_1 = \cos^{-1}(Tf_1 \cdot \nu)$. Then there exists $r_1 \in (0, \delta_1)$ such that $f_1(r_1, 0) < -1$.

Now set $\eta_2 = -(f_1(r_1, 0) + 1) > 0$ and $\delta_2 = r_1$. From Lemma 1, there exists $f_2 \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ which satisfies (1) such that $\sup_{\Omega_{\delta_2}} |f_2 - f_1| < \eta_2$, $\left|\gamma_2 - \frac{\pi}{2}\right| \le \alpha + \epsilon$ and $\lim_{\Omega \ni (x,y) \to \mathcal{O}} f_2(x, y) = \infty$, where $\gamma_2 = \cos^{-1}(Tf_2 \cdot \nu)$. Then there exists $r_2 \in (0, \delta_2)$ such that $f_2(r_2, 0) > 1$. Since $(r_1, 0) \in \Omega_{\delta_2}$,

$$f_1(r_1, 0) + 1 < f_2(r_1, 0) - f_1(r_1, 0) < -(f_1(r_1, 0) + 1)$$

and so $f_2(r_1, 0) < -1$.

Next set $\eta_3 = \min\{-(f_2(r_1,0)+1), f_2(r_2,0)-1\} > 0$ and $\delta_3 = r_2$. From Lemma 1, there exists $f_3 \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ which satisfies (1) such that $\sup_{\Omega_{\delta_3}} |f_3 - f_2| < \eta_3$, $\left|\gamma_3 - \frac{\pi}{2}\right| \le \alpha + \epsilon$ and $\lim_{\Omega \ni (x,y) \to \mathcal{O}} f_3(x,y) = -\infty$, where $\gamma_3 = \cos^{-1}(Tf_3 \cdot \nu)$. Then there exists $r_3 \in (0, \delta_3)$ such that $f_3(r_3,0) < -1$. Since $(r_1,0), (r_2,0) \in \Omega_{\delta_2}$, we have

$$f_2(r_1,0)+1 < f_3(r_1,0) - f_2(r_1,0) < -(f_2(r_1,0)+1)$$

and

$$-(f_2(r_2,0)-1) < f_3(r_2,0) - f_2(r_2,0) < f_2(r_2,0)-1;$$

hence $f_3(r_1,0) < -1$ and $1 < f_3(r_2,0)$.

Continuing to define f_n and r_n inductively, we set

$$\eta_{n+1} = \min_{1 \le j \le n} |f_n(r_j,0) - (-1)^j| \quad \text{and} \quad \delta_{n+1} = \min\{r_n, \tfrac{1}{n}\}.$$

From Lemma 1, there exists $f_{n+1} \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \{\mathcal{O}\})$ satisfying (1) such that $\sup_{\Omega_{\delta_{n+1}}} |f_{n+1} - f_n| < \eta_{n+1}$, $\left|\gamma_{n+1} - \frac{\pi}{2}\right| \le \alpha + \epsilon$ and $\lim_{\Omega \ni (x,y) \to \mathcal{O}} f_{n+1}(x,y) = (-1)^{n+1}\infty$, where $\gamma_{n+1} = \cos^{-1}(Tf_{n+1} \cdot \nu)$. Then there exists $r_{n+1} \in (0, \delta_{n+1})$ such that $(-1)^{n+1} f_{n+1}(r_{n+1},0) > 1$. For each $j \in \{1, \dots, n\}$ which is an even number, we have

$$-(f_n(r_j,0)-1) < f_{n+1}(r_j,0) - f_n(r_j,0) < f_n(r_j,0)-1$$

and so $1 < f_{n+1}(r_j,0)$. For each $j \in \{1, \dots, n\}$ which is an odd number, we have

$$f_n(r_j,0)+1 < f_{n+1}(r_j,0) - f_n(r_j,0) < -(f_n(r_j,0)+1)$$

and so $f_{n+1}(r_j,0) < -1$.

As in [Lancaster and Siegel 1996; Siegel 1980], there is a subsequence of $\{f_n\}$, still denoted $\{f_n\}$, which converges pointwise and uniformly in the C^1 norm on $\overline{\Omega}_\delta$ for each $\delta > 0$ as $n \to \infty$ to a solution $f \in C^2(\Omega) \cap C^1(\overline{\Omega} \setminus \mathcal{O})$ of (1). For each $j \in \mathbb{N}$ which is even, $f_n(r_j,0) > 1$ for each $n \in \mathbb{N}$ and so $f(r_j,0) \ge 1$. For each $j \in \mathbb{N}$ which is odd, $f_n(r_j,0) < -1$ for each $n \in \mathbb{N}$ and so $f(r_j,0) \le -1$. Therefore

$$\lim_{r \to 0^+} f(r,0) \text{ does not exist, even as an infinite limit,}$$

and so $Rf(0)$ does not exist.

Since Ω is symmetric with respect to the x-axis and $\gamma_n(x,y)$ is an even function of y, $f(x,y)$ is an even function of y. Now suppose that there exists $\theta_0 \in [-\alpha, \alpha]$ such that $Rf(\theta_0)$ exists; then $\theta_0 \ne 0$. From the symmetry of f, $Rf(-\theta_0)$ must also exist and $Rf(-\theta_0) = Rf(\theta_0)$. Set

$$\Omega' = \{(r\cos\theta, r\sin\theta) : 0 < r < \delta_0, -\theta_0 < \theta < \theta_0\} \subset \Omega.$$

Since f has continuous boundary values on $\partial\Omega'$, $f \in C^0(\overline{\Omega'})$ and so $Rf(0)$ does exist, which is a contradiction. Thus $Rf(\theta)$ does not exist for any $\theta \in [-\alpha, \alpha]$. $\square$

References

[Concus and Finn 1970] P. Concus and R. Finn, "On a class of capillary surfaces", *J. Analyse Math.* **23**:1 (1970), 65–70. MR Zbl

[Crenshaw et al. 2017] J. N. Crenshaw, A. K. Echart, and K. E. Lancaster, "A generalization of 'Existence and behavior of the radial limits of a bounded capillary surface at a corner"', 2017. To appear in *Pacific J. Math.* arXiv

[Lancaster 1989] K. E. Lancaster, "Existence and nonexistence of radial limits of minimal surfaces", *Proc. Amer. Math. Soc.* **106**:3 (1989), 757–762. MR Zbl

[Lancaster and Siegel 1996] K. E. Lancaster and D. Siegel, "Existence and behavior of the radial limits of a bounded capillary surface at a corner", *Pacific J. Math.* **176**:1 (1996), 165–194. Correction in **179**:2 (1997), 397–402. MR Zbl

[Siegel 1980] D. Siegel, "Height estimates for capillary surfaces", *Pacific J. Math.* **88**:2 (1980), 471–515. MR Zbl

Received February 1, 2017.

COLM PATRIC MITCHELL
DEPARTMENT OF MATHEMATICS, STATISTICS, AND PHYSICS
WICHITA STATE UNIVERSITY
WICHITA, KS 67260-0033
UNITED STATES
mitchell@math.wichita.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.179

INITIAL-SEED RECURSIONS AND DUALITIES FOR d-VECTORS

NATHAN READING AND SALVATORE STELLA

We present an initial-seed-mutation formula for d-vectors of cluster variables in a cluster algebra. We also give two rephrasings of this recursion: one as a duality formula for d-vectors in the style of the g-vectors/c-vectors dualities of Nakanishi and Zelevinsky, and one as a formula expressing the highest powers in the Laurent expansion of a cluster variable in terms of the d-vectors of any cluster containing it. We prove that the initial-seed-mutation recursion holds in a varied collection of cluster algebras, but not in general. We conjecture further that the formula holds *for source-sink moves on the initial seed* in an arbitrary cluster algebra, and we prove this conjecture in the case of surfaces.

1. Introduction

This paper concerns the search for an *initial-seed recursion* for $\boldsymbol{d}$-vectors: a recursive formula for how $\boldsymbol{d}$-vectors change under mutation of initial seeds. We begin this introduction by providing background on cluster algebras, seeds, and $\boldsymbol{d}$-vectors.

The origins of cluster algebras lie in the study of totally positive matrices, generalized by Lusztig [1994] to a notion of totally positive elements in any reductive group. Indeed, the recursive definition of cluster algebras extends and generalizes a recursion on minimal sets of minors whose positivity implies total positivity of matrices. Cluster algebras were introduced by Fomin and Zelevinsky [1999; 2002a], who conjectured that the coordinate ring of any *double Bruhat cell* (i.e., any intersection of two Bruhat cells for opposite Borel subgroups) is a cluster algebra.

Nathan Reading was partially supported by NSF grant DMS-1101568.
MSC2010: 13F60.
Keywords: cluster algebra, denominator vector, initial-seed recursion, marked surface.

(As it turns out, the natural choice of cluster algebra is a subring of the double Bruhat cell, proper in some cases. In general, the double Bruhat cell coincides with a related larger algebra called an *upper cluster algebra* [Berenstein et al. 2005].)

Since their introduction, cluster algebras and/or their underlying combinatorics and geometry have been found in widely different settings. Some of these settings — and some early references — are algebraic geometry (Grassmannians [Scott 2006] and tropical analogues [Speyer and Williams 2005]), discrete dynamical systems (rational recurrences [Carroll and Speyer 2004; Fomin and Zelevinsky 2002b]), higher Teichmüller theory [Fock and Goncharov 2006; 2009], PDE (KP solitons [Kodama and Williams 2011; 2014]), Poisson geometry [Gekhtman et al. 2003; 2005], representation theory of quivers/finite dimensional algebras [Buan et al. 2006; 2007; Caldero et al. 2006; Caldero and Keller 2008; Marsh et al. 2003], scattering diagrams [Gross et al. 2014; 2015; Kontsevich and Soibelman 2014], (related to mirror symmetry, Donaldson–Thomas theory, and integrable systems, and string theory), and Y-systems in thermodynamic Bethe Ansatz [Fomin and Zelevinsky 2003b].

We begin by reviewing the definition of a (coefficient free) *cluster algebra*. An *exchange matrix* $B = (b_{ij})$ is a skew-symmetrizable $n \times n$ integer matrix (meaning that there exist positive integers d_i such that $d_i b_{ij} = -d_j b_{ji}$ for every i and j). We write $\mathbb{T}_n$ for the n-regular tree with edges properly labeled $1, \dots, n$, and we distinguish one vertex t_0 as the "initial" vertex. We will write $t \overset{k}{\text{——}} t'$ to indicate that t and t' are connected by an edge labeled k. We define a function $t \mapsto B_t$ that labels each vertex of $\mathbb{T}_n$ with an exchange matrix. Specifically, we set B_{t_0} equal to some "initial" exchange matrix B_0 and, for each edge $t \overset{k}{\text{——}} t'$ with $B_t = (b_{ij})$, we insist that $B_{t'} = (b'_{ij})$ be given by

$$b'_{ij} = \begin{cases} -b_{ij} & \text{if } i = k \text{ or } j = k, \\ b_{ij} + \operatorname{sgn}(b_{kj})[b_{ik}b_{kj}]_+ & \text{otherwise.} \end{cases} \tag{1-1}$$

Here and elsewhere in the text, the notation $[a]_+$ means $\max(a, 0)$ while $\operatorname{sgn}(a)$ is the sign of a.

Taking $x_1, \dots, x_n$ to be indeterminates, we also label each vertex t of $\mathbb{T}_n$ with an n-tuple $(x_{1;t}, \dots, x_{n;t})$ of rational functions in $x_1, \dots, x_n$ called *cluster variables*. The label on t_0 consists of the indeterminates: $x_{i;t_0} = x_i$ for all i. The remaining cluster variables are prescribed by *exchange relations*. For each edge $t \overset{k}{\text{——}} t'$, we have $x_{i;t'} = x_{i;t}$ for all $i \neq k$ and

$$x_{k;t}x_{k;t'} = \prod_{i=1}^{n} x_{i;t}^{[b_{ik}]_+} + \prod_{i=1}^{n} x_{i;t}^{[-b_{ik}]_+}, \tag{1-2}$$

where the b_{ik} are entries of B_t.

Each pair $(B_t, (x_{1;t}, \dots, x_{n;t}))$ is called a *seed*. When t and t' are connected by an edge $t \overset{k}{\text{——}} t'$, the relationship between the seeds $(B_t, (x_{1;t}, \dots, x_{n;t}))$ and $(B_{t'}, (x_{1;t'}, \dots, x_{n;t'}))$ is called *mutation in direction* k. The (coefficient-free) cluster algebra $\mathscr{A}(B_0)$ associated to the initial exchange matrix B_0 is the algebra (a subalgebra of the field of rational functions in $x_1, \dots, x_n$) generated by the set

$$\{x_{i;t} : t \in \mathbb{T}_n, i = 1, \dots, n\}$$

of all cluster variables. Typically, there are infinitely many cluster variables; when the set $\{x_{i;t} : t \in \mathbb{T}_n, i = 1, \dots, n\}$ is finite, we say that B_0 is of *finite type*.

The first fundamental result on cluster algebras is the *Laurent phenomenon* [Fomin and Zelevinsky 2002a, Theorem 3.1]. The exchange relations define the cluster variables as rational functions in $x_1, \dots, x_n$. The Laurent phenomenon is the assertion that each cluster variable is in fact a Laurent polynomial (a polynomial divided by a monomial). This implies in particular that each cluster variable has a *denominator vector* or $\boldsymbol{d}$*-vector*. The $\boldsymbol{d}$-vector of $x_{i;t}$ is a vector $\boldsymbol{d}_{j;t}$ with n entries, whose j-th entry is the power of x_j^{-1} that appears as a factor of $x_{i;t}$. In principle, the $\boldsymbol{d}$-vector may have negative entries (when powers of x_j appear in the numerator of $x_{i;t}$), but in practice this only happens when $x_{i;t}$ equals some x_j.

Denominator vectors are fundamental to the theory of cluster algebras in many ways, and they are also significant in other settings beginning with Fomin and Zeleivinsky's proof [2003b] of Zamolodchikov's periodicity conjecture on Y-systems in the theory of thermodynamic Bethe ansatz. They are also important in representation theory. Each skew-symmetric $n \times n$ exchange matrix B defines a *quiver* (i.e., a directed graph) Q on the vertices $1, \dots, n$. (The signs of entries give the direction of arrows and the magnitudes of entries give multiplicities of arrows.) In the case where B is skew-symmetric and acyclic, the $\boldsymbol{d}$-vectors of cluster variables are exactly the dimension vectors of *rigid* indecomposable modules over the path algebra of Q (modules with no self-extensions). (See [Buan et al. 2007; Caldero et al. 2006].) In combinatorics, the $\boldsymbol{d}$-vectors, realized as *almost positive roots* in an associated root system, are central to the structure of *generalized associahedra* and thus play a role in Coxeter–Catalan combinatorics [Armstrong 2009; Fomin and Reading 2007] and are interesting in more general settings such as subword complexes, multiassociahedra, graph associahedra, and so forth.

Once we know the Laurent phenomenon, the exchange relations (1-2) imply a recursion on $\boldsymbol{d}$-vectors $\boldsymbol{d}_{j;t}$, given later as (2-4). This recursion is a "final-seed recursion" because it describes how $\boldsymbol{d}$-vectors (computed with respect to a fixed *initial* seed) change when we mutate the *final* seed $(B_t, (x_{1;t}, \dots, x_{n;t}))$.

We are now prepared to discuss the search for an initial-seed recursion for $\boldsymbol{d}$-vectors, describing how $\boldsymbol{d}$-vectors at a fixed *final* seed change under mutation of *initial* seeds. It is widely expected (see, for example, [Fomin and Zelevinsky 2007, Remark 7.7]) that no satisfactory initial-seed-mutation recursion holds in general, and indeed we do not produce one. However, a very nice initial-seed-mutation recursion holds in a varied collection of cluster algebras (including the case considered in [Fomin and Zelevinsky 2007, Remark 7.7]). This recursion turns out to be equivalent to a beautiful duality formula in the style of the $\boldsymbol{g}$-vectors/$\boldsymbol{c}$-vectors dualities of Nakanishi and Zelevinsky [Nakanishi 2011; Nakanishi and Zelevinsky 2012].

The first thing one notices when looking for such a recursion is that, to understand how denominators change when the initial seed is mutated, one must know something about a related family of integer vectors. Specifically, if $(x_1, \dots, x_n)$ is the initial cluster, then the *negation* of the $\boldsymbol{d}$-vector of a cluster variable x is the vector of lowest powers of the x_i occurring in the expression for x as a Laurent polynomial in $x_1, \dots, x_n$. We define the $\boldsymbol{m}$*-vector* of x to be the vector of *highest* powers of the x_i occurring in x. Our initial-seed-mutation recursion for $\boldsymbol{d}$-vectors is equivalent to a description of the $\boldsymbol{m}$-vectors in a given cluster in terms of the $\boldsymbol{d}$-vectors in the same cluster.

In many cases, one can establish the three formulas (2-1)–(2-3) by reading off the duality directly from expressions for denominator vectors found in the literature [Ceballos and Pilaud 2015; Fomin et al. 2008; Lee et al. 2014]. In particular, all of them hold in finite type, in rank two (i.e., $n=2$), and more intriguingly, in nontrivial examples arising from marked surfaces.

We conjecture that the initial-seed-mutation recursion holds in the case of source-sink moves in arbitrary cluster algebras. We prove this conjecture for cluster algebras arising from surfaces. Dylan Rupel and the second author [2017] proved the conjecture in the case where B is acyclic, using a categorification of quantum cluster algebras.

Besides their usefulness in understanding denominator vectors, the $\boldsymbol{m}$-vectors may be of independent interest. A major goal in the study of cluster algebras is to give explicit formulas for the cluster variables. Work in this direction includes realizing cluster variables as “lambda lengths” in the surfaces case [Fomin and Thurston 2012], combinatorial formulas in rank two [Lee and Schiffler 2013], in some finite types [Musiker 2011; Schiffler 2008], and for some surfaces [Musiker and Schiffler 2010; Musiker et al. 2011; Schiffler and Thomas 2009], interpretations in terms of the representation theory of quivers, beginning with [Caldero and Chapoton 2006], and formulas in terms of “broken lines” in scattering diagrams [Gross et al. 2014]. Short of a complete description of a cluster variable, one might instead describe its Newton polytope (the convex hull of the exponent vectors of the Laurent monomials occurring in its Laurent expansion). However, as far as the authors are aware, there are no general results describing Newton polytopes. (For a description in one finite-type case, see [Kalman 2014].)

Together, the $\boldsymbol{d}$-vectors and $\boldsymbol{m}$-vectors amount to coarse information about

Newton polytopes, namely their "bounding boxes." Given a polytope P in $\mathbb{R}^n$, define the *tight bounding box* of P to be the smallest box $[a_1, b_1] \times \cdots \times [a_n, b_n]$ containing P. (Readers who pay attention to bounding boxes of graphics files will find the notion familiar.) Equivalently, for each $i = 1, \ldots, n$, the values a_i and b_i are respectively the minimum and maximum of the i-th coordinates of points in P. It is convenient to describe the tight bounding box by specifying the vectors $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$. The tight bounding box of the Newton polytope of a Laurent polynomial f in $x_1, \ldots, x_n$ is $[a_1, b_1] \times \cdots \times [a_n, b_n]$ such that a_i is the lowest power of x_i occurring in any Laurent monomial of f, and b_i is the highest power of x_i occurring. Thus when x is a cluster variable written as a Laurent polynomial in the initial cluster $(x_1, \ldots, x_n)$, the tight bounding box of the Newton polytope of x is given by the negation of the $\boldsymbol{d}$-vector and by the $\boldsymbol{m}$-vector.

2. Results

Our notation is in the spirit of [Fomin and Zelevinsky 2007] and [Nakanishi and Zelevinsky 2012]. As before, the notation $[a]_+$ means $\max(a, 0)$. We will apply the operators max, $|\cdot|$, and $[\cdot]_+$ entry-wise to vectors and matrices. We continue to write $\mathbb{T}_n$ for the n-regular tree with edges properly labeled $1, \ldots, n$. Symbols like t, t_0, t', etc. will stand for vertices of $\mathbb{T}_n$. The notation $t \overset{k}{\text{———}} t'$ indicates an edge in $\mathbb{T}_n$ labeled k. In what follows, the initial seed is allowed to vary, so we need to be able to indicate the initial seed as part of the notation. Thus, the notation $B_t^{B_0;t_0}$ stands for the exchange matrix at t, where B_0 is the exchange matrix at t_0. Similarly, $x_{j;t}^{B_0;t_0}$ stands for the (coefficient-free) cluster variable indexed by j in the (labeled) seed at t, and $\boldsymbol{d}_{j;t}^{B_0;t_0}$ is the denominator vector of $x_{j;t}^{B_0;t_0}$ with respect to the cluster at t_0.

Given a matrix A, let $A^{\bullet k}$ be the matrix obtained from A by replacing all entries outside the k-th column with zeros. Similarly, $A^{k\bullet}$ is obtained by replacing entries outside the k-th row with zeros. Let J_k be the matrix obtained from the identity matrix by replacing the kk-entry by -1. The superscript T stands for transpose.

We fix $(x_1, \ldots, x_n)$ to be the initial cluster (the cluster at t_0). We write $D_t^{B_0;t_0}$ for the matrix whose j-th *column* is $\boldsymbol{d}_{j;t}^{B_0;t_0}$ and $D_{ij;t}^{B_0;t_0}$ for the ij-entry of that matrix. Each $x_{j;t}^{B_0;t_0}$ is a Laurent polynomial in $x_1, \ldots, x_n$. (This is the Laurent phenomenon, [Fomin and Zelevinsky 2002a, Theorem 3.1].) Let $M_t^{B_0;t_0}$ be the matrix whose ij-entry $M_{ij;t}^{B_0;t_0}$ is the maximum, over all of the (Laurent) monomials in $x_{j;t}^{B_0;t_0}$, of the power of x_i occurring in the monomial. Write $\boldsymbol{m}_{j;t}^{B_0;t_0}$ for the j-th *column* of $M_t^{B_0;t_0}$ and call this the j-th $\boldsymbol{m}$*-vector* at t.

We now present a duality property for denominator vectors that holds in some cluster algebras, as well as two equivalent properties: an initial-seed-mutation recursion for denominator vectors and a formula for the M-matrix at a given seed in terms of the D-matrix at the same seed.

Property D ($D =$ matrix duality). *For vertices $t_0, t \in \mathbb{T}_n$, writing B_t as shorthand for $B_t^{B_0;t_0}$,*

$$(D_t^{B_0;t_0})^T = D_{t_0}^{(-B_t)^T;t}. \tag{2-1}$$

Property R (initial-seed-mutation recursion for D-matrices). *Suppose $t_0 \overset{k}{\text{——}} t_1$ is an edge in $\mathbb{T}_n$ and write B_1 for $\mu_k(B_0)$. Then*

$$D_t^{B_1;t_1} = J_k D_t^{B_0;t_0} + \max\bigl([B_0^{k\bullet}]_+ D_t^{B_0;t_0}, [-B_0^{k\bullet}]_+ D_t^{B_0;t_0}\bigr). \tag{2-2}$$

The recursion in Property R is not on individual denominator vectors, but rather on an entire cluster of denominator vectors. For $i \neq k$, the i-th entry of each denominator vector is unchanged, while row k of the D-matrix (the vector of k-th entries in denominator vectors) transforms by a recursion similar to the usual recursion (2-4) below for how denominator vectors change under mutation.

Property M (M-matrices in terms of D-matrices). *For vertices $t_0, t \in \mathbb{T}_n$,*

$$M_t^{B_0;t_0} = -D_t^{B_0;t_0} + \max\bigl([B_0]_+ D_t^{B_0;t_0}, [-B_0]_+ D_t^{B_0;t_0}\bigr). \tag{2-3}$$

When Property M holds, in particular, the entire tight bounding box of a cluster variable x can be determined directly from the denominator vectors of any cluster containing x.

Our first main result is the following theorem, which we prove in Section 3.

Theorem 2.1. *Fix a (coefficient-free) cluster pattern $t \mapsto (B_t^{B_0;t_0}, (x_{1;t}, \ldots, x_{n;t}))$. The following are equivalent:*

(1) *Property D holds for all t_0 and t.*

(2) *Property R holds for all t_0, t, and k.*

(3) *Property M holds for all t_0 and t.*

A natural question is to characterize the cluster algebras in which Properties D, R, and M hold. As a start towards answering this question, we prove the following three theorems in Section 4. In every case, the proof is to read off Property D using a known formula for the denominator vectors.

Theorem 2.2. *Properties D, R, and M holds in any cluster pattern whose exchange matrices are 2×2.*

Theorem 2.3. *Properties D, R, and M hold in any cluster pattern of finite type.*

Theorem 2.4. *Properties D, R, and M hold for a cluster algebra arising from a marked surface if and only if the marked surface is one of the following:*

(1) *A disk with at most one puncture* (*finite types A and D*).

(2) *An annulus with no punctures and one or two marked points on each boundary component* (*affine types* $\tilde{A}_{1,1}$, $\tilde{A}_{2,1}$, *and* $\tilde{A}_{2,2}$).

(3) *A disk with two punctures and one or two marked points on the boundary component* (*affine types* $\tilde{D}_3$ *and* $\tilde{D}_4$).

(4) *A sphere with four punctures and no boundary components.*

(5) *A torus with exactly one marked point* (*either one puncture or one boundary component containing one marked point*).

In Section 3, we also prove some easier relations on D-matrices and M-matrices that hold in general. The first of these shows that, to understand how D-matrices transform under mutation of the initial seed, one must understand M-matrices.

Proposition 2.5. *Suppose* $t_0 \overset{k}{\text{———}} t_1$ *is an edge in* $\mathbb{T}_n$. *Then* $D_t^{B_1;t_1}$ *is obtained by replacing the* k*-th row of* $D_t^{B_0;t_0}$ *with the* k*-th row of* $M_t^{B_0;t_0}$. *That is,*

$$D_t^{B_1;t_1} = D_t^{B_0;t_0} - (D_t^{B_0;t_0})^{k\bullet} + (M_t^{B_0;t_0})^{k\bullet}.$$

The final-seed mutation recursion on denominator vectors [Fomin and Zelevinsky 2007, (7.6)–(7.7)] is given in matrix form as follows. The initial D-matrix $D_{t_0}^{B_0;t_0}$ is the negative of the identity matrix, and for each edge $t \overset{k}{\text{———}} t'$ in $\mathbb{T}_n$,

$$D_{t'}^{B_0;t_0} = D_t^{B_0;t_0} J_k + \max\bigl(D_t^{B_0;t_0}[(B_t^{B_0;t_0})^{\bullet k}]_+,\, D_t^{B_0;t_0}[(-B_t^{B_0;t_0})^{\bullet k}]_+\bigr). \tag{2-4}$$

Note that neither product of matrices inside the max in (2-4) has any nonzero entry outside the k-th column. It turns out that $\boldsymbol{m}$-vectors satisfy the same recursion, but with different initial conditions.

Proposition 2.6. *The initial* M*-matrix* $M_{t_0}^{B_0;t_0}$ *is the identity matrix. Given an edge* $t \overset{k}{\text{———}} t'$ *in* $\mathbb{T}_n$,

$$M_{t'}^{B_0;t_0} = M_t^{B_0;t_0} J_k + \max\bigl(M_t^{B_0;t_0}[(B_t^{B_0;t_0})^{\bullet k}]_+,\, M_t^{B_0;t_0}[(-B_t^{B_0;t_0})^{\bullet k}]_+\bigr).$$

Finally, we present some conjectures and results on Property R in the context of source-sink moves. Suppose that in the exchange matrix B_0, all entries in row k weakly agree in sign. That is, either all entries in row k are nonnegative (and equivalently all entries in column k are nonpositive) or all entries in row k are nonpositive (and equivalently all entries in column k are nonnegative). In this case, mutation of B_0 in direction k is often called a *source-sink move*, referring to the operation on quivers of reversing all arrows at a source or a sink. We conjecture that Property R holds when mutation at k is a source-sink move. In this case, Equation (2-2) has a particularly simple form.

Conjecture 2.7. *Suppose $t_0 \overset{k}{\text{———}} t_1$ is an edge in $\mathbb{T}_n$ and B_1 is $\mu_k(B_0)$. If all entries in row k of B_0 weakly agree in sign, then*

$$D_t^{B_1;t_1} = J_k D_t^{B_0;t_0} + \big[|B_0^{k\bullet}| D_t^{B_0;t_0}\big]_+ \tag{2-5}$$

We also make two other closely related conjectures. Let A be the *Cartan companion* of B_0, defined by setting $A_{ii} = 2$ for all i and $A_{ij} = -|(B_0)_{ij}|$ for $i \neq j$. Then A is a (generalized) Cartan matrix and thus defines a root system and a root lattice in the usual way. It also defines a (generalized) Weyl group W, generated by simple reflections $s_1, \dots, s_n$ given by $s_k(\alpha_\ell) = \alpha_\ell - A_{k\ell}\alpha_k$, where the α_i are the simple roots. If β is in the root lattice, then write $[\beta : \alpha_i]$ for the coefficient of α_i in the simple root coordinates of β. Then $[s_k(\beta) : \alpha_i] = [\beta : \alpha_i]$ if $i \neq k$ and $[s_k(\beta) : \alpha_k] = -[\beta : \alpha_k] + \sum_{\ell=1}^{n} |(B_0)_{k\ell}|[\beta : \alpha_\ell]$. Following [Fomin and Zelevinsky 2003b, Section 2], we define a piecewise linear modification σ_k of s_k by setting $[\sigma_k(\beta) : \alpha_i] = [\beta : \alpha_i]$ if $i \neq k$ and $[\sigma_k(\beta) : \alpha_k] = -[\beta : \alpha_k] + \sum_{\ell=1}^{n} |(B_0)_{k\ell}|\big[[\beta : \alpha_\ell]\big]_+$. We think of σ_k as a map on (certain) integer vectors by interpreting them as simple root coordinates of vectors in the root lattice. We also think of σ_k as a map on integer matrices by applying it to each *column*.

Conjecture 2.8. *Suppose $t_0 \overset{k}{\text{———}} t_1$ is an edge in $\mathbb{T}_n$ and B_1 is $\mu_k(B_0)$. If all entries in row k of B_0 weakly agree in sign, then $D_t^{B_1;t_1} = \sigma_k D_t^{B_0;t_0}$.*

To relate Conjecture 2.8 to Conjecture 2.7, we quote the following conjecture, which is a significant weakening of [Fomin and Zelevinsky 2007, Conjecture 7.4]. We will say a matrix D *has signed columns* if every column of D either has all nonnegative entries or all nonpositive entries. Similarly, D *has signed rows* if every row of D either has all nonnegative entries or all nonpositive entries.

Conjecture 2.9. *For all $t \in \mathbb{T}_n$, the matrix $D_t^{B_0;t_0}$ has signed columns.*

Conjecture 2.9 is not the same as another weakening of [Fomin and Zelevinsky 2007, Conjecture 7.4], namely "sign-coherence of $\boldsymbol{d}$-vectors," which asserts that for all $t \in \mathbb{T}_n$, the matrix $D_t^{B_0;t_0}$ has signed *rows*.

We prove the following easy proposition in Section 3.

Proposition 2.10. *If Conjecture 2.9 holds, then Conjectures 2.7 and 2.8 are equivalent.*

Theorems 2.2 and 2.3 imply Conjecture 2.7 in the rank-two and finite-type cases, and Theorem 2.4 implies it for certain surfaces. Rupel and Stella [2017] proved Conjectures 2.8 and 2.9 (and thus Conjecture 2.7) for B acyclic. As further evidence in support of the conjectures in general, we prove the following theorem in Section 5.

Theorem 2.11. *Conjectures 2.7 and 2.8 hold in cluster algebras arising from marked surfaces.*

3. Proofs of general results

We begin with the proof of Proposition 2.6, followed by the proof of Proposition 2.5. To make the proof of Proposition 2.6 completely clear, we point out two lemmas about highest powers in multivariate (Laurent) polynomials. Both are completely obvious when looked at in the right way, but otherwise one might convince oneself to worry. Given a Laurent polynomial p, we write $m_i(p)$ for the highest power of x_i occurring in a term of p.

Lemma 3.1. *Given Laurent polynomials f and g in $x_1, \dots, x_n$, we have $m_i(fg) = m_i(f) + m_i(g)$.*

Proof. Write $f = f_a x_i^a + f_{a+1} x_i^{a+1} + \cdots + f_k x_i^k$ and $g = g_b x_i^b + g_{b+1} x_i^{b+1} + \cdots + g_\ell x_i^\ell$ such that the f_j and g_j are polynomials in the variables besides x_i and f_k and g_ℓ are nonzero. Then the highest power of x_i in fg is $k+\ell$. (Otherwise f_k and g_ℓ are zero divisors.) □

Lemma 3.2. *Suppose p is a Laurent polynomial over $\mathbb{C}$ in $x_1, \dots, x_n$ and f and g are polynomials in $\mathbb{C}[x_1, \dots, x_n]$ such that $f/g = p$. Then $m_i(p) = m_i(f) - m_i(g)$.*

Proof. Since p is a Laurent polynomial, we can factor f as $a \cdot c$ and g as $b \cdot c$ such that b is a monomial. It is immediate that $m_i(p) = m_i(a) - m_i(b)$. Applying Lemma 3.1, we have $m_i(f) - m_i(g) = m_i(a) + m_i(c) - m_i(b) - m_i(c) = m_i(p)$. □

Proof of Proposition 2.6. Throughout this proof, we omit superscripts $B_0; t_0$. The first assertion of the proposition is trivial. To establish the second assertion, we compute $M_{ij;t'}$, the highest power of x_i occurring in $x_{j;t'}$, in terms of M_t. If $j \neq k$, then $x_{j;t'} = x_{j;t}$, so $M_{ij;t'} = M_{ij;t}$ as given in the proposition. If $j = k$, then the exchange relation [Fomin and Zelevinsky 2007, (2.8)], with trivial coefficients, is

$$x_{k;t'} = (x_{k;t})^{-1} \left(\prod_\ell (x_{\ell,t})^{[B_{\ell k;t}]_+} + \prod_\ell (x_{\ell,t})^{[-B_{\ell k;t}]_+} \right). \tag{3-1}$$

Write U for the expression $\prod_\ell (x_{\ell,t})^{[B_{\ell k;t}]_+} + \prod_\ell (x_{\ell,t})^{[-B_{\ell k;t}]_+}$. Each factor $x_{\ell;t}$ in U has a subtraction-free expression: an expression as a ratio of two polynomials in $x_1, \dots, x_n$ with nonnegative coefficients. Therefore each term in U has a subtraction-free expression. Write the first term as a/c and the second term as b/d, where a, b, c, and d are polynomials with nonnegative coefficients. The sum U is then $\frac{ad}{cd} + \frac{bc}{cd}$. Since all of these expressions are subtraction-free, there is no cancellation, so $m_i(U) = m_i\left(\frac{ad}{cd} + \frac{bc}{cd}\right) = \max\left(m_i\left(\frac{ad}{cd}\right), m_i\left(\frac{bc}{cd}\right)\right)$, which equals $\max\left(m_i\left(\frac{a}{c}\right), m_i\left(\frac{b}{d}\right)\right)$ which in turn equals

$$\max\left(m_i\left(\prod_\ell (x_{\ell,t})^{[B_{\ell k;t}]_+} \right), m_i\left(\prod_\ell (x_{\ell,t})^{[-B_{\ell k;t}]_+} \right) \right).$$

Returning now to expressions for the $x_{\ell;t}$ as Laurent polynomials, Lemma 3.1 lets us conclude that $m_i(U) = \max\bigl(\sum_\ell M_{i\ell;t}[B_{\ell k;t}]_+, \sum_\ell M_{i\ell;t}[-B_{\ell k;t}]_+\bigr)$.

Now, writing $x_{k;t}$ as a rational function p/q with $m_i(p)-m_i(q)=m_i(x_{k;t})$ and writing U as a rational function r/s with $m_i(r)-m_i(s)=m_i(U)$, Equation (3-1) lets us write $x_{k;t'}$ as $\frac{qr}{ps}$, so Lemmas 3.1 and 3.2 imply that $M_{ik;t'} = m_i(q)-m_i(p)+m_i(r)-m_i(s) = -M_{ik;t} + \max\bigl(\sum_\ell M_{i\ell;t}[B_{\ell k;t}]_+, \sum_\ell M_{i\ell;t}[-B_{\ell k;t}]_+\bigr)$ as desired. $\square$

Proof of Proposition 2.5. The cluster at t_1 is obtained from $(x_1, \dots, x_n)$ by removing x_k and replacing it with a new cluster variable x'_k. The two are related by

$$x_k = (x'_k)^{-1}\Bigl(\prod_\ell x_\ell^{[b_{\ell k}]_+} + \prod_\ell x_\ell^{[-b_{\ell k}]_+}\Bigr), \tag{3-2}$$

where the $b_{\ell k}$ are entries of B_0.

To show that the k-th row of $D_t^{B_1;t_1}$ equals the k-th row of $M_t^{B_0;t_0}$, we appeal to the Laurent phenomenon to write the cluster variable $x_{j;t}^{B_0;t_0}$ in the form

$$\frac{N(x_1, \dots, x_n)}{\prod_i x_i^{D_{ij;t}^{B_0;t_0}}},$$

for some polynomial N not divisible by any of the x_i. We write $N = N_0 + N_1 x_k + \cdots + N_p x_k^p$, where the N_q are polynomials not involving x_k, with $N_p \neq 0$. Then (3-2) lets us write $x_{j;t}^{B_0;t_0}$ as

$$\frac{N_0 + N_1 \frac{\left(\prod_\ell x_\ell^{[b_{\ell k}]_+} + \prod_\ell x_\ell^{[-b_{\ell k}]_+}\right)}{x'_k} + \cdots + N_p \frac{\left(\prod_\ell x_\ell^{[b_{\ell k}]_+} + \prod_\ell x_\ell^{[-b_{\ell k}]_+}\right)^p}{(x'_k)^p}}{x_1^{D_{1j;t}^{B_0;t_0}} \cdots \left(\frac{\left(\prod_\ell x_\ell^{[b_{\ell k}]_+} + \prod_\ell x_\ell^{[-b_{\ell k}]_+}\right)}{x'_k}\right)^{D_{kj;t}^{B_0;t_0}} \cdots x_n^{D_{nj;t}^{B_0;t_0}}} \tag{3-3}$$

The numerator of (3-3) can be factored as $(x'_k)^{-p}$ times a polynomial not divisible by x'_k. The denominator can be factored as $(x'_k)^{-D_{kj;t}^{B_0;t_0}}$ times a polynomial not involving x'_k. We conclude that $D_{kj;t}^{B_1,t_1}$ is $-D_{kj;t}^{B_0;t_0} + p$. The latter equals $M_{kj;t}^{B_0;t_0}$.

To show that $D_t^{B_1;t_1}$ agrees with $D_t^{B_0;t_0}$ outside of row k, we fix $i \neq k$ and consider a subtraction-free expression for $x_{j;t}^{B_0;t_0}$. The Laurent phenomenon implies that this expression can be simplified to a Laurent polynomial. The simplification can, if one wishes, be done in two stages, by first factoring out all powers of x_i from the rational expression and then canceling the other factors. After the first stage, we have written $x_{j;t}^{B_0;t_0}$ as $x_i^{-D_{ij;t}^{B_0;t_0}} \cdot \frac{f}{g}$ where f and g are subtraction-free polynomials not divisible by x_i. Replacing x_k in this expression by the right side of (3-2), we find that no additional powers of x_i can be extracted. (Since the right side of (3-2) is also subtraction-free, we obtain a new subtraction-free expression. In particular, there can be no cancellation, so a power of x_i can be extracted if and only if it is a factor in every term of the numerator or a factor in every term of the denominator. But the right side of (3-2) is not divisible by any nonzero power of x_i.) We conclude that $D^{B_1,t_1}_{ij;t} = D^{B_0;t_0}_{ij;t}$. $\square$

We next prove Theorem 2.1. Specifically, the theorem follows from the next three propositions, which more carefully specify the relations among the three properties.

Proposition 3.3. *For a fixed choice of B_0, t_0, t and k, let t_1 be the vertex of $\mathbb{T}_n$ such that $t_0 \overset{k}{\frac{\quad}{}} t_1$ and write B_1 for $\mu_k(B_0)$. Suppose (2-1) holds at B_0, t_0, t and also at B_1, t_1, t. Then (2-2) holds for the same B_0, t_0, t, k.*

Proof. We apply (2-1) at B_1, t_1, t, then (2-4), then (2-1) at B_0, t_0, t.

$$\begin{aligned} D_t^{B_1;t_1} &= (D_{t_1}^{(-B_t)^T;t})^T \\ &= \big(D_{t_0}^{(-B_t)^T;t} J_k + \max\big(D_{t_0}^{(-B_t)^T;t}[(B_0^T)^{\bullet k}]_+,\, D_{t_0}^{(-B_t)^T;t}[(-B_0^T)^{\bullet k}]_+\big)\big)^T \\ &= J_k (D_{t_0}^{(-B_t)^T;t})^T + \max\big([B_0^{k\bullet}]_+ (D_{t_0}^{(-B_t)^T;t})^T,\, [-B_0^{k\bullet}]_+ (D_{t_0}^{(-B_t)^T;t})^T\big) \\ &= J_k D_t^{B_0;t_0} + \max\big([B_0^{k\bullet}]_+ D_t^{B_0;t_0},\, [-B_0^{k\bullet}]_+ D_t^{B_0;t_0}\big) \end{aligned}$$

In the second line, we use the fact that $B_{t_0}^{(-B_t)^T;t} = -B_0^T$. $\square$

Proposition 3.4. *Fix a (coefficient-free) cluster pattern $t \mapsto (B_t, (x_{1;t}, \dots, x_{n;t}))$ and vertices t_0 and t of $\mathbb{T}_n$, connected by edges*

$$t_0 \overset{k_1=k}{\frac{\qquad}{}} t_1 \overset{k_2}{\frac{\qquad}{}} \cdots \overset{k_m}{\frac{\qquad}{}} t_m = t.$$

Suppose that, for all $i = 1, \dots, m$, equation (2-2) holds for the edge $t_{i-1} \overset{k_i}{\frac{\quad}{}} t_i$. Then

$$(D_t^{B_{t_0};t_0})^T = D_{t_0}^{(-B_t)^T;t}.$$

Proof. We argue by induction on m. For $m = 0$ (i.e., $t = t_0$), (2-1) says that the negative of the identity matrix is symmetric. Equation (2-2) is symmetric in switching t_0 and t_1, because $B_0^{k\bullet} = -B_1^{k\bullet}$. Thus for $m > 0$, we can use (2-2) for the edge $t_1 \overset{k_1}{\frac{\quad}{}} t_0$ to write

$$D_t^{B_0;t_0} = J_k D_t^{B_1;t_1} + \max\big([B_1^{k\bullet}]_+ D_t^{B_1;t_1},\, [-B_1^{k\bullet}]_+ D_t^{B_1;t_1}\big) \tag{3-4}$$

By induction, we rewrite the right side of (3-4) as

$$\begin{aligned} &J_k (D_{t_1}^{(-B_t)^T;t})^T + \max\big([B_1^{k\bullet}]_+ (D_{t_1}^{(-B_t)^T;t})^T,\, [-B_1^{k\bullet}]_+ (D_{t_1}^{(-B_t)^T;t})^T\big) \\ &\qquad = \big(D_{t_1}^{(-B_t)^T;t} J_k + \max\big(D_{t_1}^{(-B_t)^T;t}[(B_1^T)^{\bullet k}]_+,\, D_{t_1}^{(-B_t)^T;t}[(-B_1^T)^{\bullet k}]_+\big)\big)^T. \end{aligned}$$

By (2-4), this is $(D_{t_0}^{(-B_t)^T;t})^T$. $\square$

Proposition 3.5. *For a fixed choice of B_0, t_0, t and k, (2-2) holds if and only if (2-3) holds* in the k-th row.

Proof. Equation (2-3) holds in the k-th row if and only if

$$(M_t^{B_0;t_0})^{k\bullet} = (-D_t^{B_0;t_0})^{k\bullet} + \max\bigl([B_0^{k\bullet}]_+ D_t^{B_0;t_0}, [-B_0^{k\bullet}]_+ D_t^{B_0;t_0}\bigr).$$

This equation is equivalent to (2-2) in light of Proposition 2.5. □

This completes the proof of Theorem 2.1.

To conclude this section, we establish Proposition 2.10 by proving a more detailed statement. Recall that a matrix D has signed columns if every column of D either has all nonnegative entries or all nonpositive entries.

Proposition 3.6. *Suppose $t_0 \xrightarrow{k} t_1$ is an edge in $\mathbb{T}_n$ and B_1 is $\mu_k(B_0)$. If $D_t^{B_0;t_0}$ has signed columns, then the right side of (2-5) equals $\sigma_k D_t^{B_0;t_0}$.*

Proof. Let β be the vector in the root lattice with simple root coordinates $\boldsymbol{d}_{j;t}^{B_0;t_0}$. For $i \neq k$, the ij-entry of the right side of (2-5) is $[\beta : \alpha_i]$. The kj-entry of the right side of (2-5) is $-[\beta : \alpha_k] + \bigl[\sum_{\ell=1}^n |(B_0)_{k\ell}|[\beta : \alpha_\ell]\bigr]_+$. By hypothesis, all of the simple root coordinates of β weakly agree in sign, so $\bigl[\sum_{\ell=1}^n |(B_0)_{k\ell}|[\beta : \alpha_\ell]\bigr]_+$ is $\sum_{\ell=1}^n |(B_0)_{k\ell}|\bigl[[\beta : \alpha_\ell]\bigr]_+$. Thus the right side of (2-5) is $\sigma_k\beta$. □

4. Duality and recursion in certain cluster algebras

We now prove Theorems 2.2, 2.3, and 2.4.

4A. ***Rank two.*** The proof of Theorem 2.2 uses a formula for rank-two denominator vectors due to Lee, Li, and Zelevinsky [2014, (1.13)].

Proof of Theorem 2.2. The 2-regular tree $\mathbb{T}_n$ is an infinite path. We label its vertices t_k for $k \in \mathbb{Z}$, and abbreviate B_{t_k} by B_k. As the situation is very symmetric, it is enough to take $B_0 = \left[\begin{smallmatrix} 0 & b \\ -c & 0 \end{smallmatrix}\right]$ with b and c nonnegative and establish (2-1) for $t = t_k$ with $k \geq 0$. When $bc < 4$, the cluster pattern is of finite type and (2-1) can be checked easily (if a bit tediously) by hand. Alternatively, one can appeal to Theorem 2.3, which we prove below. For $bc \geq 4$, the denominator vectors are given by [Lee et al. 2014, (1.13)]. Equation (2-1) is easy when $k = 1$, so we assume $k \geq 2$. The labeled cluster associated to the vertex t_k is $\{x_{k+1}, x_{k+2}\}$ if k is even and $\{x_{k+2}, x_{k+1}\}$ if k is odd.

If k is even, we use [Lee et al. 2014, (1.13)] to write

$$D_{t_k}^{B_0;t_0} = \begin{bmatrix} S_{\frac{k-2}{2}}(u) + S_{\frac{k-4}{2}}(u) & bS_{\frac{k-2}{2}}(u) \\ cS_{\frac{k-4}{2}}(u) & S_{\frac{k-2}{2}}(u) + S_{\frac{k-4}{2}}(u) \end{bmatrix} \tag{4-1}$$

where $u = bc - 2$ and the S_p are Chebyshev polynomials of the second kind. (In fact, here we do not need to know anything about the S_p except that they are functions of u.) We can similarly use [Lee et al. 2014, (1.13)] to write an expression for $D_{t_0}^{-B_k^T;t_k}$. Since k is even $B_k = B_0$, and thus $-B_k^T = -B_0^T = \left[\begin{smallmatrix} 0 & c \\ -b & 0 \end{smallmatrix}\right]$. To apply [loc. cit., (1.13)] in this case, we must switch the role of b and c. When we do so, keeping in mind that we move now in the negative direction, we obtain exactly the transpose of the right side of (4-1).

If k is odd, we obtain

$$D_{t_k}^{B_0;t_0} = \begin{bmatrix} S_{\frac{k-1}{2}}(u) + S_{\frac{k-3}{2}}(u) & bS_{\frac{k-3}{2}}(u) \\ cS_{\frac{k-3}{2}}(u) & S_{\frac{k-3}{2}}(u) + S_{\frac{k-5}{2}}(u) \end{bmatrix} \tag{4-2}$$

In this case, $B_k = -B_0$, so $-B_k^T = B_0^T = \left[\begin{smallmatrix} 0 & -c \\ b & 0 \end{smallmatrix}\right]$. Noticing that $-B_k^T$ is obtained from B_0 by simultaneously swapping the rows and the columns, when we use [loc. cit., (1.13)] to write an expression for $D_{t_0}^{-B_k^T;t_k}$, we also swap the rows and columns. The result is exactly the transpose of the right side of (4-2). $\square$

4B. ***Finite type.*** The proof of Theorem 2.3 uses a result of Ceballos and Pilaud [2015] giving denominator vectors in finite type, with respect to any initial seed, in terms of the compatibility degrees defined at any acyclic seed. In [Fomin and Zelevinsky 2003a], it is shown that in every cluster pattern of finite type, there exists an exchange matrix B_0 that is bipartite and whose Cartan companion A is of finite type. The cluster variables appearing in the cluster pattern are in bijection with the almost positive roots in the root system for A. Given an almost positive root β, we will write $x(\beta)$ for the corresponding cluster variable. There is a *compatibility degree* $(\alpha, \beta) \mapsto (\alpha \parallel \beta) \in \mathbb{Z}_{\geq 0}$ defined on almost positive roots encoding some of the combinatorial properties of the cluster algebra. In particular two cluster variables $x(\alpha)$ and $x(\beta)$ belong to the same cluster if and only if the roots α and β are *compatible* (i.e., if their compatibility degree is zero). Maximal sets of compatible roots are called *(combinatorial) clusters* and they correspond to the *(algebraic) clusters* in the cluster algebra. In the same paper Fomin and Zelevinsky also showed that compatibility degrees encode denominator vectors with respect to the bipartite initial seed.

Ceballos and Pilaud extended this dramatically in the following result. (We follow them in modifying the definition of compatibility degree in an inconsequential way in order to make it easier to state the theorem. Specifically, we take $(\alpha \parallel \alpha) = -1$ rather than $(\alpha \parallel \alpha) = 0$.)

Theorem 4.1 [Ceballos and Pilaud 2015, Corollary 3.2]. *Let $\{\beta_1, \dots, \beta_n\}$ be a cluster and let γ be an almost positive root. Then the $\mathbf{d}$-vector of $x(\gamma)$ with respect to the cluster $\{x(\beta_1), \dots, x(\beta_n)\}$ is given by $[(\beta_1 \parallel \gamma), \dots, (\beta_n \parallel \gamma)]$.*

Since B_0 is skew-symmetrizable, passing from B_0 to $-B_0^T$ has the effect of preserving the signs of entries while transposing the Cartan companion A. The almost positive roots for A^T are the almost positive coroots associated to A. The following is [Fomin and Zelevinsky 2003b, Proposition 3.3(1)].

Proposition 4.2. *If α and β are almost positive roots and $\alpha^\vee$ and $\beta^\vee$ are the corresponding coroots, then* $(\alpha \parallel \beta) = (\beta^\vee \parallel \alpha^\vee)$.

Proof of Theorem 2.3. The cluster pattern assigns some algebraic cluster to t_0 and some algebraic cluster to t, and each of the algebraic clusters is encoded by some combinatorial cluster. Let $\{\beta_1, \dots, \beta_n\}$ be the combinatorial cluster at t_0 and let $\{\gamma_1, \dots, \gamma_n\}$ be the combinatorial cluster at t. Now Theorem 4.1 and Proposition 4.2 are exactly Property D at t_0 and t. □

4C. *Marked surfaces.* The proof of Theorem 2.4 relies on a result of Fomin, Shapiro, and Thurston [2008, Theorem 8.6] giving denominator vectors in terms of tagged arcs. We will assume familiarity with the basic definitions of cluster algebras arising from marked surfaces.

Recall that tagged arcs are in bijection with cluster variables and tagged triangulations are in bijection with clusters, except in the case of once-punctured surfaces with no boundary components, where plain-tagged arcs are in bijection with cluster variables and plain-tagged triangulations are in bijection with clusters. We write $\alpha \mapsto x(\alpha)$ for this bijection. Given tagged arcs α and β, there is an *intersection number* $(\alpha|\beta)$ such that the following theorem [Fomin et al. 2008, Theorem 8.6] holds.

Theorem 4.3. *Given tagged arcs α and β and a cluster $(x_1, \dots, x_n)$ with $x_i = x(\alpha)$, the i-th component of the denominator vector of $x(\beta)$ with respect to the cluster $(x_1, \dots, x_n)$ is $(\alpha|\beta)$.*

In an exchange pattern arising from a marked surface, every exchange matrix B_t is skew-symmetric, so $(-B_t)^T = B_t$. Thus we have the following corollary to Theorem 4.3.

Corollary 4.4. *In an exchange pattern arising from a marked surface, Property D holds if and only if the intersection number is symmetric (i.e., $(\alpha|\beta) = (\beta|\alpha)$ on all tagged arcs α and β that correspond to cluster variables).*

The intersection number $(\alpha|\beta)$ is defined in [Fomin et al. 2008, Definition 8.4] to be the sum of four quantities A, B, C, and D. To define these, we choose α_0 and β_0 to be non-self-intersecting curves homotopic (relative to the set of marked points) to α and β, and intersecting with each other the minimum possible number of times, transversally each time. The quantity A is the number of intersection points of α_0 and β_0 (excluding intersections at their endpoints). The quantity B is zero unless α_0 is a loop (i.e., unless the two endpoints of α_0 coincide). If α_0 is a loop, let a be its endpoint. We number the intersections as $b_1, \dots, b_k$ in the order they are encountered when following β_0 in some direction. For each $i = 1, \dots, k-1$, there is a unique segment $[a, b_i]$ of α_0 having endpoints a and b_i and not containing b_{i+1}. There is also a unique segment $[a, b_{i+1}]$ of α_0 having endpoints a and b_{i+1} and not containing b_i. Let $[b_i, b_{i+1}]$ be the segment of β_0 connecting b_i to b_{i+1}.

The quantity B is $\sum_{i=1}^{k-1} B_i$, where B_i is -1 if the segments $[a, b_i]$, $[a, b_{i+1}]$, and $[b_i, b_{i+1}]$ define a triangle that is contractible and $B_i = 0$ otherwise. The quantity C is zero unless α_0 and β_0 are equal up to isotopy relative to the set of marked points, in which case $C = -1$. The quantity D is the number of ends of β that are incident to an endpoint of α and carry, at that endpoint, a different tag from the tag of α at that endpoint.

The quantities A and C are patently symmetric in α and β, so we need not consider them in this section. It is pointed out in [Fomin et al. 2008, Example 8.5] that D can fail to be symmetric. The quantity B can also fail to be symmetric. Examples will occur below.

Some immediate observations will be helpful.

Observation 1. In a surface having no tagged arcs that are loops, B is always 0 and D is also always symmetric.

Observation 2. In a surface having no punctures, D is always zero.

Observation 3. In a surface having exactly one puncture and no boundary components, D is always zero on tagged arcs corresponding to cluster variables.

For the second observation, recall that notched tagging may occur only at punctures. For the third observation, recall that in a surface with exactly one puncture and no boundary components, tagged arcs correspond to cluster variables if and only if they are tagged plain.

To prove one direction of Theorem 2.4, we show that $B + D$ is symmetric in the cases listed in the theorem. First, recall that a tagged arc may not bound a once-punctured monogon and may not be homotopic to a segment of the boundary between two adjacent marked points. In particular, there are no loops in a disc with at most one puncture, in the unpunctured annulus with 2 marked points, in the twice-punctured disk with one marked point on its boundary, or in the four-times-punctured sphere. Thus Observation 1 shows that $B + D$ is symmetric in the cases described in (1) and (4), and in the simplest cases described in (2) and (3).

In the remaining cases described in (2), Observation 2 shows that D is always zero. We are interested in pairs of arcs containing at least one loop (otherwise B is zero in both directions). Because an arc may not be homotopic to a boundary segment, there are no loops at a marked point if it is the only marked point on its boundary component. A marked point that is not the only marked point on its boundary component supports exactly one loop. If there are two marked points on one component and one marked point on the other, we are in the situation of Figure 1, left. In this case, numbering the points as in the figure, the only two loops in the surface are based one at 1 and one at 2. The remaining arcs start at 3, spiral around some number of times and then reach either 1 or 2. The only arc that intersects one of the loops more than once is the other loop. These have

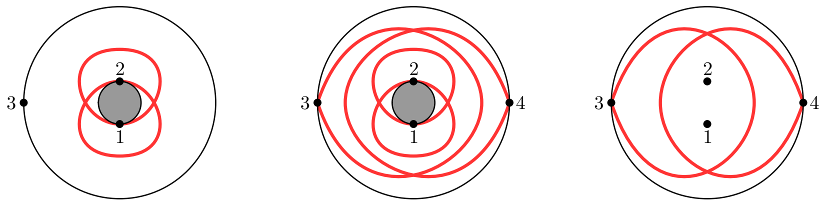


Figure 1

$B = -1$ in both directions, so B is symmetric in this case. If there are two marked points on each boundary component, the argument is similar and only slightly more complicated. There are four loops, as illustrated in Figure 1, center. Each of the remaining arcs connects a point of one boundary to a point of the other boundary, with some number of spirals. Again, for any of the four loops there is only one other arc intersecting it more than once; it is the loop based at the other marked point on the same boundary component. For each pair of intersecting loops we calculate $B = -1$ in both directions. We have finished case (2).

The remaining case (a disk with two punctures and two boundary points) in (3) is similar to the cases in (2). There is a loop at each marked point on the boundary but no other loop, as illustrated in Figure 1, right. In particular, the tagging at these loops is plain, and we see that D is symmetric. There is exactly one arc connecting the two boundary points and four tagged arcs (all with the same underlying arc) connecting the two punctures. The remaining arcs have a boundary point at one endpoint and spiral around the punctures some number of times before ending at one of the punctures, with either tagging there. Once again, the only arc that intersects more than once one of the loops is the other loop, and we again have $B = -1$ in both directions.

The remaining two cases are described in (5). We first consider the once-punctured torus. In this cases, $D = 0$ by Observation 3, so it remains to show that B is symmetric. We will show that in fact B is zero on all pairs of arcs. Arcs in the once-punctured torus are well-known to be in bijection with rational slopes, including the infinite slope. (See, for example, [Reading 2014, Section 4].) Each such slope can be written uniquely as a reduced fraction $\frac{b}{a}$ such that $a \geq 0$ and that $b = 1$ whenever $a = 0$. If we take the universal cover (the plane $\mathbb{R}^2$) of the torus mapping each integer point to the puncture, the arc indexed by a slope $\frac{b}{a}$ lifts to a straight line segment connecting the origin to the point (a, b). (The same arc also lifts to all integer translates of that line segment.)

It is now easy to see that $B = 0$ for arcs in the once-punctured torus. For any two arcs α and β, let α_0 and β_0 be the curves on the torus obtained by projecting the associated straight line segments in the plane. This choice of representatives minimizes the number of intersections as can be seen by looking at the universal

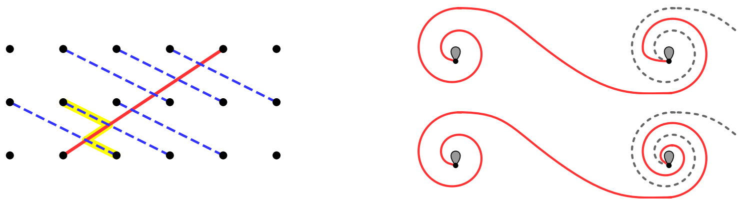


Figure 2

cover. Let a and $b_1, \dots, b_k$ be the points as in the definition of B. Given some i between 1 and $k-1$, concatenate the curves $[a, b_i]$, $[b_i, b_{i+1}]$, and $[b_{i+1}, a]$, and consider the lift of the concatenated curve to the plane. This lifted curve consists of two parallel line segments and one line segment not parallel to the other two. In particular, it is impossible for the lifted curve to start and end at the same point. The situation is illustrated in Figure 2, left, where a lift of β_0 is shown as a solid line, several lifts of α_0 are shown as dotted lines and a lift of the three concatenated curves is highlighted. By the standard argument on fundamental groups and universal covers, we see that the concatenation of $[a, b_i]$, $[b_i, b_{i+1}]$, and $[b_{i+1}, a]$ is not a contractible triangle, and we conclude that $B = 0$ on α and β.

The final case for this direction of the proof is the torus with one boundary component and one marked point. In this case, D is again zero, this time by Observation 2, so we will show that B is symmetric. We think of the boundary component as a "fat point" on the torus. With this trick, we can again consider lifts of arcs to the plane. Each arc lifts to a curve connecting the origin to an integer point (a, b) with a and b satisfying the same conditions as above for the once-punctured torus. However, for each such (a, b), there is a countable collection of arcs connecting the origin to (a, b). Specifically, for each integer k, the arc may wind k times clockwise about the fat origin point before going to (a, b). (Negative values of k specify counterclockwise spirals.) Since (a, b) and the origin both project to the same fat point on the torus, the number and direction of spirals at (a, b) is determined almost uniquely by k. There are two possibilities for each k, illustrated in Figure 2, right, for the case where $(a, b) = (1, 0)$.

For each arc α, choosing the right change of basis of the integer lattice, we may as well assume that the lift of α connects the origin to the point $(1, 0)$. Furthermore, there is a homeomorphism from the torus to itself that rotates the fat point and changes the number of spirals of α at the origin and at $(1, 0)$. Rotating a half-integer number of full turns, we can assume α lifts to a straight horizontal line segment from $(0, 0)$ to $(1, 0)$. Possibly reflecting the plane through the horizontal line containing the origin (to offset the effect of a half-turn), we can assume that α looks like the solid arc shown in Figure 3, with the boundary component above the origin in the picture.

Now take another arc β and consider a lift of β connecting the origin to (a, b).

Figure 3

Since another lift connects $(-a, -b)$ to the origin, we may as well take $b \geq 0$. Up to a reflection in a vertical line, we can assume that the lift of β spirals clockwise (if it spirals at all) as it leaves the origin. Fixing one possible number of spirals of β at $(0, 0)$ and fixing some (a, b) with $b > 0$, the two possibilities for the lift of β are shown as dashed arcs in Figure 3. Nonzero contributions to B can arise only from segments that remain close to the fat point: by the same argument as for the once-punctured torus, the segments that do not stay near the fat point contribute nothing. Therefore it is enough to analyze the intersections of α and β near the origin. In each of the two possibilities we highlight in Figure 3 the segments $[a_i, a_{i+1}]$ of α and $[b_j, b_{j+1}]$ of β giving nonzero contributions. In the pictured examples, B is symmetric in α and β. It is easy to see that the symmetry survives when the number of spirals changes. The case $b = 0$ looks slightly different, but B is still symmetric for essentially the same reasons. (Look back, for example at Figure 2, right.)

We have proved one direction of Theorem 2.4. To prove the other direction, we need to show that $B + D$ fails to be symmetric in certain cases. In each case, the failure of symmetry can be illustrated in a figure. Here, we list the cases and indicate, for each case, the corresponding figure. In some cases, we also include some comments in italics. In each case, α is the solid arc and β is the dashed arc; they intersect in at most two points. We omit the labeling a, a_1, a_2, b, b_1, b_2 not to clutter the pictures. This will complete the proof of Theorem 2.4.

(a) *A surface with genus greater than 1* (Figure 4, left). We show the genus-2 case. Pairs of edges in the octahedron are identified as indicated by the numbering and the arrows. Since all taggings are plain, $D = 0$. However, B is asymmetric (-1 in one direction and 0 in the other). The marked point shown in the figure is a puncture, but the same example works with the marked point on a boundary component. For higher genus or to have additional punctures, one can start with the surface shown and perform a connected sum, cutting a disk from the interior of the octagon shown.

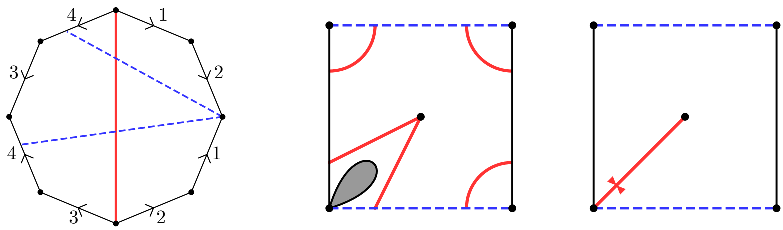


Figure 4

(b) *A torus with 2 or more marked points* (Figure 4, right). Opposite pairs of edges in the square are identified. If the marked point at the corners of the square is on a boundary component, then the arcs shown in the left picture of the figure have $D = 0$ but B is asymmetric (taking values 0 and -1). Additional punctures and/or boundary components may exist, but the arcs α and β can always be chosen so that the triangle $[b, a_1]$,$[a_1, a_2]$, $[a_2, b]$ is contractible while the triangle $[a, b_1]$, $[b_1, b_2]$, $[b_2, a]$ is not. If the marked point at the corners is a puncture, then the right picture applies. In this case, $B = 0$ but D is asymmetric because one of the arcs is a loop and the other is not.

(c) *A sphere with 3 or more boundary components and possibly some punctures* (Figure 5, left). We show a disk with 2 additional boundary components. For the arcs shown, $D = 0$ but B is asymmetric. Again, additional punctures and/or boundary components may exist, but the triangle $[a, b_1]$, $[b_1, b_2]$, $[b_2, a]$ is contractible.

(d) *An annulus with one or more punctures* (Figure 5, center). $B = 0$ and D is asymmetric on the arcs shown.

(e) *An unpunctured annulus with 3 or more marked points on one of its boundary components* (Figure 5, right). B is asymmetric and $D = 0$.

(f) *A disk with 3 or more punctures* (Figure 6, left). $B = 0$ and D is asymmetric.

(g) *A disk with 2 punctures and 3 or more marked points on the boundary* (Figure 6, center). B is asymmetric and $D = 0$.

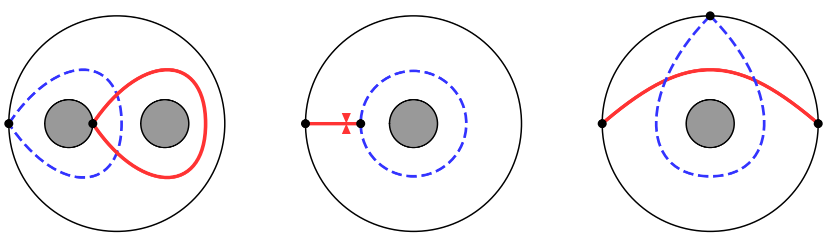


Figure 5

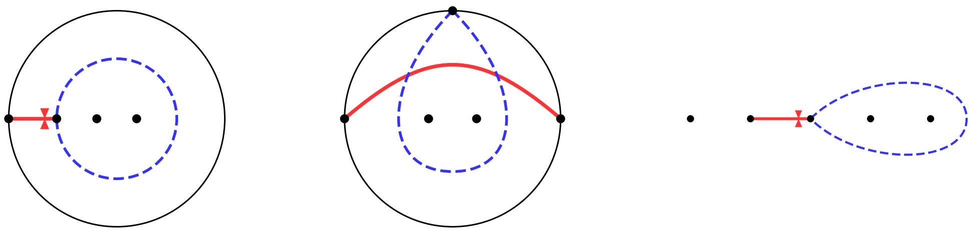


Figure 6

(h) *A sphere with 5 or more punctures* (Figure 6, right). We show a local patch of the sphere containing all of the punctures. $B = 0$ and D is asymmetric.

5. Source-sink moves on triangulated surfaces

In this section, we prove Theorem 2.11, the assertion that Conjectures 2.7 and 2.8 hold for marked surfaces. Conjecture 2.9 holds for surfaces because the stronger conjecture [Fomin and Zelevinsky 2007, Conjecture 7.4] for surfaces is an easy consequence of [Fomin et al. 2008, Theorem 8.6]. Thus by Proposition 2.10, we need only to prove the assertion about Conjecture 2.7. In light of Theorem 4.3, the task is to prove a certain identity on intersection numbers. This identity is already known (as a special case of Property R) for the surfaces listed in Theorem 2.4, and it will be convenient in what follows that we need not consider those surfaces.

Suppose α is a tagged arc in a tagged triangulation T and suppose α' is the arc obtained by flipping α in T. We may as well take T to be obtained from an ideal triangulation T° by applying the map τ of [Fomin et al. 2008, Definition 7.2] to each arc. (Any other tagged triangulation could be obtained from such a triangulation by changing tags, which by definition [loc. cit., Definition 9.6] does not affect the associated B-matrix.) In particular, $B(T) = B(T^\circ)$. We will abuse notation and denote by the same Greek letters both ideal arcs and their corresponding tagged arcs. Suppose all of the entries in the row of $B(T)$ indexed by α weakly agree in sign. Because of the symmetry between $D_t^{B_0;t_0}$ and $D_t^{B_1;t_1}$ in (2-5), we may as well assume that all entries in the row of $B(T)$ indexed by α are nonnegative; in this case we will say that "α is a source" alluding to the usual encoding of skew-symmetric exchange matrices by quivers.

Let β be any other arc. Keeping in mind that (2-5), like (2-2) before it, is true outside of row k by Proposition 2.5, the task is to prove the following identity:

$$(\alpha'|\beta) = -(\alpha|\beta) + \sum_{\gamma \in T} b_{\alpha\gamma}(\gamma|\beta) \tag{5-1}$$

where $b_{\alpha\gamma}$ is the entry of $B(T)$ in the row indexed by α and column indexed by γ.

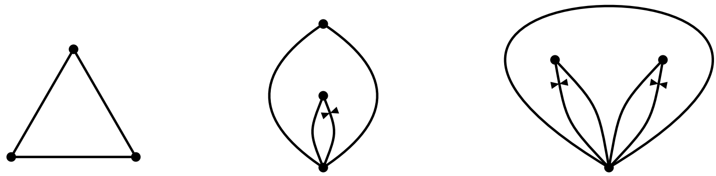


Figure 7. Tagged puzzle pieces.

The key observation in our proof is that the entries $b_{\alpha\gamma}$ depends only on how T looks locally near α. Therefore we begin our analysis by constructing a short list of possible local configurations. To do this we build the surface and the ideal triangulation T simultaneously by adjoining *puzzle pieces* as in [Fomin et al. 2008, Section 4]. There the *ideal* triangulation T° is built from puzzle pieces, but to save a step, we apply the map τ to the puzzle pieces *before* assembling, rather than *after*. The resulting *tagged puzzle pieces* are shown in Figure 7. We will refer to them (from left to right in the Figure) as triangle pieces, digon pieces, and monogon pieces. The external edges of digon pieces are distinguishable (up to reversing the orientation of the surface) and we will call them the left edge and the right edge according to how they are pictured in Figure 7. Similarly, the two pairs of internal arcs in a monogon piece are distinguishable, and we will call them the left pair and right pair according to Figure 7.

Puzzle pieces are joined by gluing along their outer edges. Unjoined outer edges become part of the boundary of the surface. In [loc. cit.], one specific triangulation is mentioned that cannot be obtained from these puzzle pieces, but it is a triangulation of the 4-times punctured sphere, so by Theorem 2.4, we need not consider it.

The list of possible local configurations around α, given α is a source, appears in Figure 8. (We leave out the cases where Theorem 2.4 applies.) In the figure, areas just outside the boundary are marked in gray. The curve α is labeled, or if two curves might be a source, both of them are labeled α.

To obtain this list, recall that the entries in the row indexed by α are determined by the triangles of T° containing α or, if α is the folded side of a self-folded triangle, by the triangles containing the other side of that self-folded triangle. (See [Fomin et al. 2008, Definition 4.1].) In particular, if α is an internal arc in a digon or

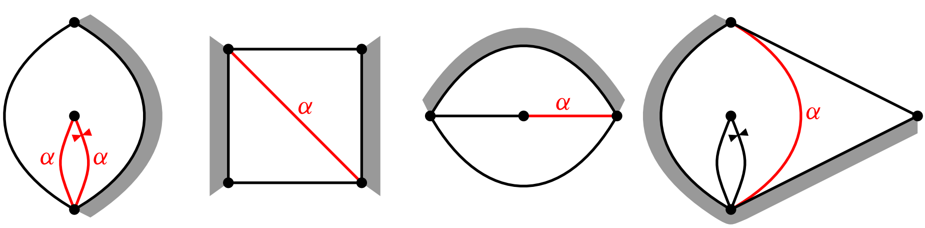


Figure 8. Possible local configurations surrounding a source.

monogon piece, the entries in the row indexed by α are determined completely within the piece. Both internal arcs in the digon piece are sources if and only if the right external edge of the digon is on the boundary, as shown in the first (i.e., leftmost) picture of Figure 8. We need not consider the case where both external edges of the digon piece are on the boundary, because Theorem 2.4 applies to a once-punctured digon.

In the monogon piece, both the arcs in the left pair are never sources, and the arcs of the right pair are sources if and only if the external edge of the monogon is on the boundary. However, we don't need to consider that case because the surface is a twice-punctured monogon, and Theorem 2.4 applies.

If α is the external edge of a monogon piece, then each of the two left internal arcs γ has $b_{\alpha\gamma} = -1$, so α is not a source. It remains, then, to consider how external edges of triangle and digon pieces can be sources. We need to consider two cases.

Suppose α is an edge in a triangle piece and suppose γ is the edge reached from α by traversing the boundary of the triangle in a counterclockwise direction. If γ is not on the boundary, then the triangle contributes -1 to $b_{\alpha\gamma}$, so α cannot be a source unless either γ is on the boundary or α and γ are also in a second triangle that contributes 1 to $b_{\alpha\gamma}$.

Next suppose α is an external edge in a digon piece. If α is the left edge, then each of the two internal arcs γ has $b_{\alpha\gamma} = -1$, so α is not a source. If α is the right edge, let γ be the left edge. As in the triangle case, α cannot be a source unless either γ is on the boundary or α and γ are also in a second triangle that contributes 1 to $b_{\alpha\gamma}$.

Putting all these observations together, we see that we must consider three more possibilities obtained by gluing a triangle or digon piece to another triangle or digon piece. We can glue two triangle pieces together along one edge with opposite edges of the resulting quadrilateral on the boundary as shown in the second picture in Figure 8. Conceivably the top and bottom arcs shown in the picture are identified, but we need not consider this case because then the surface is an annulus with two marked points on each boundary component, and Theorem 2.4 applies. We can glue two triangle pieces together along two edges, with one of the remaining edges on the boundary as shown in the third picture in Figure 8. We can glue a triangle piece along one of its edges to the right edge of a digon piece, with both the left digon edge and the triangle edge counterclockwise from the glued edge on the boundary, as shown in the fourth and last picture in Figure 8. We can glue a triangle piece along two of its edges to the two edges of a digon piece, with the remaining edge of the triangle on the boundary, but we need not consider this case, because the surface is a twice-punctured monogon, and Theorem 2.4 applies. We can glue two digon pieces, right edge to right edge, with the remaining two edges on the boundary. However, we need not consider this case either, because the surface is a

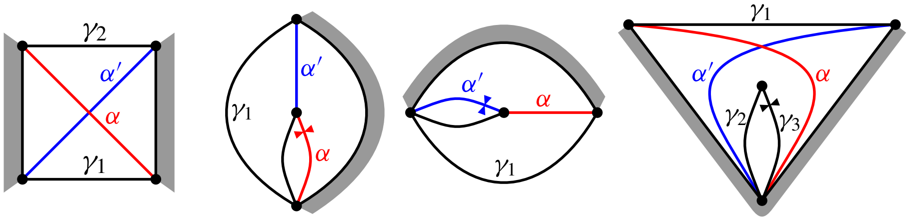


Figure 9. Possible local configurations, with more information.

twice-punctured digon and Theorem 2.4 applies. Finally, we can glue both edges of a digon piece to both edges of another digon piece, but in this case, we obtain a 3-times punctured sphere, which is explicitly disallowed in the definition of marked surfaces [Fomin et al. 2008, Definition 2.1]. Thus the four configurations in Figure 8 are the only local configurations near arcs that are sources, except in surfaces to which Theorem 2.4 applies. We will see that the first and third configurations shown are essentially equivalent for our purposes.

We observe that $(\alpha|\beta)$ is invariant under changing all taggings of α and of β at some puncture. Thus for the first (leftmost) picture in Figure 8, we may as well take α to be the arc tagged notched at the puncture. Figure 9 shows the configurations of Figure 8 with some additional information. First, the arc α', obtained by flipping α, is shown and labeled. Also, the arcs γ such that $b_{\alpha\gamma} > 0$ are labeled. There is either one arc γ_1, two arcs γ_1 and γ_2, or three arcs γ_1, γ_2 and γ_3. The pictures in Figure 9 are reordered in the order we will consider them. We have also redrawn the last configuration more symmetrically.

Recall from Section 4C that $(\alpha|\beta)$ is the sum of four quantities A, B, C, and D. As before, α_0 and β_0 are non-self-intersecting curves homotopic (relative to the set of marked points) to α and β respectively, intersecting with each other the minimum possible number of times, transversally each time. Recall that $B = 0$ unless α_0 is a loop. In the configurations of Figure 8, α_0 is never a loop. Furthermore, the quantity $b_{\alpha\gamma}$ is nonzero only if γ is in a triangle with α, and none of the arcs making triangles with α is a loop in the configurations of Figure 8. Therefore, we can ignore B in all the calculations of intersection numbers in this section. Recall also that A is the number of intersection points of α_0 and β_0 (excluding intersections at their endpoints), that $C = 0$ unless α_0 and β_0 coincide, in which case $C = -1$, and that D is the number of ends of β that are incident to an endpoint of α and carry, at that endpoint, a different tag from the tag of α at that endpoint.

Our task is simplified by several symmetries. We have already used the symmetry of changing taggings at a puncture. Also, any symmetry of a configuration that fixed α and α' or switches α and α' preserves (2-5). If the symmetry is orientation-reversing, the absolute value operation in (2-5) is crucial to the symmetry. (This absolute value has been omitted in (5-1) because we took α to be a source, not a sink.)

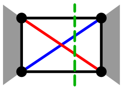
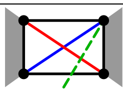
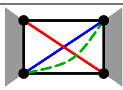
	$(\alpha\|\beta)$	$(\alpha'\|\beta)$	$(\gamma_1\|\beta)$	$(\gamma_2\|\beta)$
	$1+0+0$	$1+0+0$	$1+0+0$	$1+0+0$
	$1+0+0$	$0+0+0$	$1+0+0$	$0+0+0$
	$1+0+0$	$0+(-1)+0$	$0+0+0$	$0+0+0$

Table 1

We first consider the left picture in Figure 9. Since each marked point is on the boundary, there are no relevant taggings. Contributions to $(\alpha|\beta)$, $(\alpha'|\beta)$, $(\gamma_1|\beta)$, and $(\gamma_2|\beta)$ occur only when β intersects the interior of the quadrilateral. While β may intersect the interior of the quadrilateral a number of times, each intersection can be treated separately. In such an intersection, β may either pass through the quadrilateral, terminate at a vertex of the quadrilateral, or connect two vertices of the quadrilateral. Up to symmetry, as discussed above, there are only three possibilities. (The relevant symmetry group is the order-4 dihedral symmetry group of the rectangle shown.) Table 1 shows the possible intersections of β (shown as a dotted line) with the quadrilateral, along with the contributions to $(\alpha|\beta)$, $(\alpha'|\beta)$, $(\gamma_1|\beta)$, and $(\gamma_2|\beta)$. Each of these is given in the form $A+C+D$. The quantities $b_{\alpha\gamma_1}$ and $b_{\alpha\gamma_2}$ are both 1. In every case, we see that $(\alpha|\beta) = -(\alpha'|\beta) + (\gamma_1|\beta) + (\gamma_2|\beta)$, and therefore (5-1) holds.

Notice that the second and third pictures of Figure 9 are related by a reflection

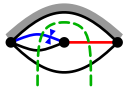
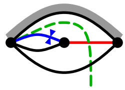
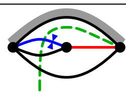
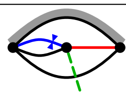
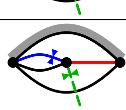
	$(\alpha\|\beta)$	$(\alpha'\|\beta)$	$(\gamma_1\|\beta)$
	$1+0+0$	$1+0+0$	$2+0+0$
	$1+0+0$	$0+0+0$	$1+0+0$
	$0+0+0$	$1+0+0$	$1+0+0$
	$0+0+0$	$0+0+1$	$1+0+0$
	$0+0+1$	$0+0+0$	$1+0+0$

Table 2

that switches α with α'. By the symmetry discussed above, we need only consider one of these configurations; we will work with the third picture. Contributions to $(\alpha|\beta)$, $(\alpha'|\beta)$, and $(\gamma_1|\beta)$ only occur when β intersects the digon, and again, we can treat each intersection separately. Table 2 shows all but four of the possible intersections of β with the configuration, and shows $(\alpha|\beta)$, $(\alpha'|\beta)$, and $(\gamma_1|\beta)$, again in the form $A+C+D$.

Since $b_{\alpha\gamma_1} = 1$, the desired relation is $(\alpha|\beta) = -(\alpha'|\beta) + (\gamma_1|\beta)$, and we see that this relation holds in every case. Not pictured in Table 2 are the four cases where β_0 coincides with α_0 or α'_0, with two possible taggings at the point in the center of the digon. In each of these cases, $(\gamma_1|\beta) = 0$ and the A terms of $(\alpha|\beta)$ and $(\alpha'|\beta)$ are both zero. The other terms are also zero, except that one of $(\alpha|\beta)$ and $(\alpha'|\beta)$ has $C = -1$ and one of $(\alpha|\beta)$ and $(\alpha'|\beta)$ has $D = 1$.

Finally, we consider the last picture in Figure 9. The two cases where β_0 coincides with α_0 or α'_0 are handled analogously to the last case of the quadrilateral condition. Up to symmetry, there are six remaining cases, pictured in Table 3. Once again, $b_{\alpha\gamma_i} = 1$ for $i \in \{1, 2, 3\}$, and the desired relation holds in every case:

$$(\alpha|\beta) = -(\alpha'|\beta) + (\gamma_1|\beta) + (\gamma_2|\beta) + (\gamma_3|\beta).$$

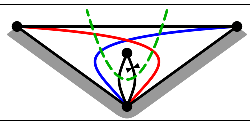
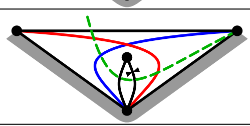
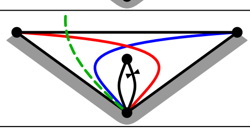
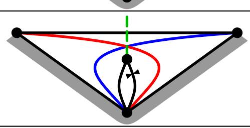
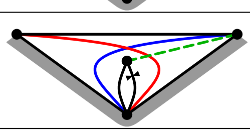
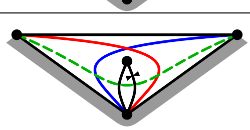
	$(\alpha\|\beta)$	$(\alpha'\|\beta)$	$(\gamma_1\|\beta)$	$(\gamma_2\|\beta)$	$(\gamma_3\|\beta)$
	2+0+0	2+0+0	2+0+0	1+0+0	1+0+0
	2+0+0	1+0+0	1+0+0	1+0+0	1+0+0
	1+0+0	0+0+0	1+0+0	0+0+0	0+0+0
	1+0+0	1+0+0	1+0+0	0+0+0	0+0+1
	1+0+0	0+0+0	0+0+0	0+0+0	0+0+1
	1+0+0	1+0+0	0+0+0	1+0+0	1+0+0

Table 3

References

[Armstrong 2009] D. Armstrong, *Generalized noncrossing partitions and combinatorics of Coxeter groups*, Mem. Amer. Math. Soc. **949**, American Mathematical Society, Providence, RI, 2009. MR Zbl

[Berenstein et al. 2005] A. Berenstein, S. Fomin, and A. Zelevinsky, "Cluster algebras, III: Upper bounds and double Bruhat cells", *Duke Math. J.* **126**:1 (2005), 1–52. MR Zbl

[Buan et al. 2006] A. B. Buan, R. Marsh, M. Reineke, I. Reiten, and G. Todorov, "Tilting theory and cluster combinatorics", *Adv. Math.* **204**:2 (2006), 572–618. MR Zbl

[Buan et al. 2007] A. B. Buan, R. J. Marsh, I. Reiten, and G. Todorov, "Clusters and seeds in acyclic cluster algebras", *Proc. Amer. Math. Soc.* **135**:10 (2007), 3049–3060. MR Zbl

[Caldero and Chapoton 2006] P. Caldero and F. Chapoton, "Cluster algebras as Hall algebras of quiver representations", *Comment. Math. Helv.* **81**:3 (2006), 595–616. MR Zbl

[Caldero and Keller 2008] P. Caldero and B. Keller, "From triangulated categories to cluster algebras", *Invent. Math.* **172**:1 (2008), 169–211. MR Zbl

[Caldero et al. 2006] P. Caldero, F. Chapoton, and R. Schiffler, "Quivers with relations and cluster tilted algebras", *Algebr. Represent. Theory* **9**:4 (2006), 359–376. MR Zbl

[Carroll and Speyer 2004] G. D. Carroll and D. Speyer, "The cube recurrence", *Electron. J. Combin.* **11**:1 (2004), art. id. 73. MR Zbl

[Ceballos and Pilaud 2015] C. Ceballos and V. Pilaud, "Denominator vectors and compatibility degrees in cluster algebras of finite type", *Trans. Amer. Math. Soc.* **367**:2 (2015), 1421–1439. MR Zbl

[Fock and Goncharov 2006] V. V. Fock and A. B. Goncharov, "Moduli spaces of local systems and higher Teichmüller theory", *Publ. Math. Inst. Hautes Études Sci.* 103 (2006), 1–211. MR Zbl

[Fock and Goncharov 2009] V. V. Fock and A. B. Goncharov, "Cluster ensembles, quantization and the dilogarithm", *Ann. Sci. Éc. Norm. Supér.* (4) **42**:6 (2009), 865–930. MR Zbl

[Fomin and Reading 2007] S. Fomin and N. Reading, "Root systems and generalized associahedra", pp. 63–131 in *Geometric combinatorics*, edited by E. Miller et al., IAS/Park City Math. Ser. **13**, American Mathematical Society, Providence, RI, 2007. MR Zbl

[Fomin and Thurston 2012] S. Fomin and D. Thurston, "Cluster algebras and triangulated surfaces, II: Lambda lengths", preprint, 2012. arXiv

[Fomin and Zelevinsky 1999] S. Fomin and A. Zelevinsky, "Double Bruhat cells and total positivity", *J. Amer. Math. Soc.* **12**:2 (1999), 335–380. MR Zbl

[Fomin and Zelevinsky 2002a] S. Fomin and A. Zelevinsky, "Cluster algebras, I: Foundations", *J. Amer. Math. Soc.* **15**:2 (2002), 497–529. MR Zbl

[Fomin and Zelevinsky 2002b] S. Fomin and A. Zelevinsky, "The Laurent phenomenon", *Adv. in Appl. Math.* **28**:2 (2002), 119–144. MR Zbl

[Fomin and Zelevinsky 2003a] S. Fomin and A. Zelevinsky, "Cluster algebras, II: Finite type classification", *Invent. Math.* **154**:1 (2003), 63–121. MR Zbl

[Fomin and Zelevinsky 2003b] S. Fomin and A. Zelevinsky, "Y-systems and generalized associahedra", *Ann. of Math.* (2) **158**:3 (2003), 977–1018. MR Zbl

[Fomin and Zelevinsky 2007] S. Fomin and A. Zelevinsky, "Cluster algebras, IV: Coefficients", *Compos. Math.* **143**:1 (2007), 112–164. MR Zbl

[Fomin et al. 2008] S. Fomin, M. Shapiro, and D. Thurston, "Cluster algebras and triangulated surfaces, I: Cluster complexes", *Acta Math.* **201**:1 (2008), 83–146. MR Zbl

[Gekhtman et al. 2003] M. Gekhtman, M. Shapiro, and A. Vainshtein, "Cluster algebras and Poisson geometry", *Mosc. Math. J.* **3**:3 (2003), 899–934. MR Zbl

[Gekhtman et al. 2005] M. Gekhtman, M. Shapiro, and A. Vainshtein, "Cluster algebras and Weil–Petersson forms", *Duke Math. J.* **127**:2 (2005), 291–311. Correction in **139**:2 (2007), 407–409. MR Zbl

[Gross et al. 2014] M. Gross, P. Hacking, S. Keel, and M. Kontsevich, "Canonical bases for cluster algebras", preprint, 2014. arXiv

[Gross et al. 2015] M. Gross, P. Hacking, and S. Keel, "Birational geometry of cluster algebras", *Algebr. Geom.* **2**:2 (2015), 137–175. MR Zbl

[Kalman 2014] A. Kalman, "Newton polytopes of cluster variables of type A_n", pp. 137–147 in *26th International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC)* (Chicago, 2014), edited by L. J. Billera and I. Novik, Assoc. Discrete Math. Theor. Comput. Sci., Nancy, France, 2014. MR

[Kodama and Williams 2011] Y. Kodama and L. K. Williams, "KP solitons, total positivity, and cluster algebras", *Proc. Natl. Acad. Sci. USA* **108**:22 (2011), 8984–8989. MR Zbl

[Kodama and Williams 2014] Y. Kodama and L. Williams, "KP solitons and total positivity for the Grassmannian", *Invent. Math.* **198**:3 (2014), 637–699. MR Zbl

[Kontsevich and Soibelman 2014] M. Kontsevich and Y. Soibelman, "Wall-crossing structures in Donaldson–Thomas invariants, integrable systems and mirror symmetry", pp. 197–308 in *Homological mirror symmetry and tropical geometry*, edited by R. Castano-Bernard et al., Lect. Notes Unione Mat. Ital. **15**, Springer, Cham, 2014. MR Zbl

[Lee and Schiffler 2013] K. Lee and R. Schiffler, "A combinatorial formula for rank 2 cluster variables", *J. Algebraic Combin.* **37**:1 (2013), 67–85. MR Zbl

[Lee et al. 2014] K. Lee, L. Li, and A. Zelevinsky, "Greedy elements in rank 2 cluster algebras", *Selecta Math.* (*N.S.*) **20**:1 (2014), 57–82. MR Zbl

[Lusztig 1994] G. Lusztig, "Total positivity in reductive groups", pp. 531–568 in *Lie theory and geometry*, edited by J.-L. Brylinski et al., Progr. Math. **123**, Birkhäuser, Boston, 1994. MR Zbl

[Marsh et al. 2003] R. Marsh, M. Reineke, and A. Zelevinsky, "Generalized associahedra via quiver representations", *Trans. Amer. Math. Soc.* **355**:10 (2003), 4171–4186. MR Zbl

[Musiker 2011] G. Musiker, "A graph theoretic expansion formula for cluster algebras of classical type", *Ann. Comb.* **15**:1 (2011), 147–184. MR Zbl

[Musiker and Schiffler 2010] G. Musiker and R. Schiffler, "Cluster expansion formulas and perfect matchings", *J. Algebraic Combin.* **32**:2 (2010), 187–209. MR Zbl

[Musiker et al. 2011] G. Musiker, R. Schiffler, and L. Williams, "Positivity for cluster algebras from surfaces", *Adv. Math.* **227**:6 (2011), 2241–2308. MR Zbl

[Nakanishi 2011] T. Nakanishi, "Periodicities in cluster algebras and dilogarithm identities", pp. 407–443 in *Representations of algebras and related topics*, edited by A. Skowroński and K. Yamagata, Eur. Math. Soc., Zürich, 2011. MR Zbl

[Nakanishi and Zelevinsky 2012] T. Nakanishi and A. Zelevinsky, "On tropical dualities in cluster algebras", pp. 217–226 in *Algebraic groups and quantum groups* (Nagoya, 2010), edited by S. Ariki et al., Contemp. Math. **565**, American Mathematical Society, Providence, RI, 2012. MR Zbl

[Reading 2014] N. Reading, "Universal geometric coefficients for the once-punctured torus", *Sém. Lothar. Combin.* **71** (2014), art. id. B71e. MR Zbl

[Rupel and Stella 2017] D. Rupel and S. Stella, "Some consequences of categorification", 2017. In preparation.

[Schiffler 2008] R. Schiffler, "A cluster expansion formula (A_n case)", *Electron. J. Combin.* **15**:1 (2008), art. id. 64. MR Zbl

[Schiffler and Thomas 2009] R. Schiffler and H. Thomas, "On cluster algebras arising from unpunctured surfaces", *Int. Math. Res. Not.* **2009**:17 (2009), 3160–3189. MR Zbl

[Scott 2006] J. S. Scott, "Grassmannians and cluster algebras", *Proc. London Math. Soc.* (3) **92**:2 (2006), 345–380. MR Zbl

[Speyer and Williams 2005] D. Speyer and L. Williams, "The tropical totally positive Grassmannian", *J. Algebraic Combin.* **22**:2 (2005), 189–210. MR Zbl

Received March 30, 2016. Revised June 26, 2017.

NATHAN READING
DEPARTMENT OF MATHEMATICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NC
UNITED STATES
reading@math.ncsu.edu

SALVATORE STELLA
DEPARTMENT OF MATHEMATICS AND DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF HAIFA
HAIFA, MOUNT CARMEL
ISRAEL
stella@mat.uniroma1.it

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.207

CODIMENSIONS OF THE SPACES OF CUSP FORMS FOR SIEGEL CONGRUENCE SUBGROUPS IN DEGREE TWO

ALOK SHUKLA

We give a computational algorithm for describing the one-dimensional cusps of the Satake compactifications for the Siegel congruence subgroups in the case of degree two for arbitrary levels. As an application of the results thus obtained, we calculate the codimensions of the spaces of cusp forms in the spaces of modular forms of degree two with respect to Siegel congruence subgroups of levels not divisible by 8. We also construct a linearly independent set of Klingen–Eisenstein series with respect to the Siegel congruence subgroup of an arbitrary level.

1. Introduction

One of the most basic questions about the spaces of modular forms is to ask for the dimensions and the codimensions of the spaces of cusp forms. For the spaces of Siegel modular forms of degree two with respect to the full modular group $\mathrm{Sp}(4, \mathbb{Z})$ the answers have been well known for several decades. However, while the answers for the spaces of modular forms with respect to Siegel congruence subgroups are not so clear, several special cases have been treated in the literature. Dimensions of the spaces of cusp forms with respect to $\Gamma_0(p)$ have been computed by Hashimoto [1983] for weights $k \geq 5$. For $\Gamma_0(2)$, Ibukiyama [1991] gave the structure of the ring of Siegel modular forms of degree 2. Poor and Yuen [2007] computed the dimensions of cusp forms for weights $k = 2, 3, 4$ with respect to $\Gamma_0(p)$ in the case of a small prime p. In [Poor and Yuen 2013] Poor and Yuen described the one-dimensional and zero-dimensional cusps of the Satake compactifications for the paramodular subgroups in the degree two case and calculated the codimensions of cusp forms. More recently Böcherer and Ibukiyama [2012] have given a formula for calculating the codimensions of the spaces of cusp forms in the spaces of modular forms of degree two with respect to Siegel congruence subgroups of

The author wishes to acknowledge his immense gratitude to Ralf Schmidt for his great help during the research and preparation of the manuscript. The author also thanks Cris Poor, R. Schulze-Pillot and Tomoyoshi Ibukiyama for their helpful comments communicated to Ralf Schmidt.
MSC2010: 11F46.
Keywords: dimension formula, Siegel modular forms, Klingen–Eisenstein series with level, cusp structure, double coset decompositions.

square-free levels. In this paper we generalize their result and give a formula for the codimensions of the spaces of cusp forms in the spaces of modular forms of degree two with respect to Siegel congruence subgroups of level N with $8 \nmid N$. Another application of the results presented here will appear in a forthcoming work on Klingen–Eisenstein series. The method used to find the codimensions of the spaces of cusp forms makes use of a result from the theory of Satake compactification. The cusp structure of the Satake compactification encodes information about the codimensions of cusp forms and works of several authors indicate that it is an important object worth investigating.

2. Notation

We shall use the following notation throughout this paper unless otherwise stated. We realize the group GSp(4) as

$$\mathrm{GSp}(4) := \{g \in \mathrm{GL}(4) \mid {}^t g J g = \lambda(g) J \text{ for some } \lambda(g) \in \mathrm{GL}(1)\},$$

with $J = \begin{bmatrix} & J_1 \\ -J_1 & \end{bmatrix}$ and $J_1 = \begin{bmatrix} & 1 \\ 1 & \end{bmatrix}$. We note that this version of GSp(4) is isomorphic to the classical version of GSp(4) and we denote this isomorphism by the map $\jmath$ which interchanges the first two rows and the first two columns of any matrix. Sp(4) is the subgroup of GSp(4) consisting of matrices with multiplier $\lambda = 1$. By $Q(\mathbb{Q})$ and $P(\mathbb{Q})$ we will denote the Klingen and Siegel parabolic subgroups of $\mathrm{GSp}(4, \mathbb{Q})$, respectively, consisting of the matrices of the form

$$Q(\mathbb{Q}) = \left\{ \begin{smallmatrix} * & * & * & * \\ & * & * & * \\ & * & * & * \\ & & & * \end{smallmatrix} \,\middle|\, * \in \mathbb{Q} \right\} \quad \text{and} \quad P(\mathbb{Q}) = \left\{ \begin{smallmatrix} * & * & * & * \\ * & * & * & * \\ & & * & * \\ & & * & * \end{smallmatrix} \,\middle|\, * \in \mathbb{Q} \right\}.$$

We define

$$s_1 := \begin{bmatrix} & 1 & & \\ 1 & & & \\ & & & 1 \\ & & 1 & \end{bmatrix}, \quad s_2 := \begin{bmatrix} 1 & & & \\ & & 1 & \\ & -1 & & \\ & & & 1 \end{bmatrix}.$$

We define the Siegel congruence subgroup as

$$\Gamma_0(N) = \Gamma_0^4(N) := \left\{ \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ a & b & * & * \\ c & d & * & * \end{bmatrix} \in \mathrm{Sp}(4, \mathbb{Z}) \,\middle|\, a, b, c, d \equiv 0 \mod N \right\}.$$

We will denote by $\Gamma(N)$ the usual principal congruence subgroup of $\mathrm{Sp}(4, \mathbb{Z})$. Next we define $\Delta(\mathbb{Z}/N\mathbb{Z}) := \{g \bmod N \mid g \in \Gamma_0(N)\}$, $\Gamma_\infty(\mathbb{Z}) := Q(\mathbb{Q}) \cap \mathrm{Sp}(4, \mathbb{Z})$.

We will use $\Gamma_0^2(N) := \left\{ \begin{bmatrix} 0 & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2, \mathbb{Z}) \mid c \equiv 0 \bmod N \right\}$ to denote the Hecke congruence subgroup of $\mathrm{SL}(2, \mathbb{Z})$ and $\Gamma_\infty^2(\mathbb{Z}) := \left\{ \pm \begin{bmatrix} 1 & b \\ & 1 \end{bmatrix} \mid b \in \mathbb{Z} \right\}$. For $Z \in \mathbb{H}_2 := \{z \in M_2(\mathbb{C}) \mid {}^t z = z, \operatorname{Im} z > 0\}$, and for any $m = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathrm{Sp}(4, \mathbb{Z})$ we define $m\langle Z \rangle :=$

$(AZ+B)(CZ+D)^{-1}$, $j(m, Z) := CZ+D$ and $m\langle Z\rangle^* = \tilde{\tau}$ for $m\langle Z\rangle = \left[\begin{smallmatrix} \tilde{\tau} & \tilde{z} \\ \tilde{z} & \tilde{\tau}' \end{smallmatrix}\right]$. We will denote by $C_0(N)$ and $C_1(N)$ the number of zero and one-dimensional cusps for $\Gamma_0(N)$ respectively, i.e.,

$$C_0(N) = \#(\Gamma_0(N)\backslash \operatorname{GSp}(4, \mathbb{Q})/P(\mathbb{Q})), \quad C_1(N) = \#(\Gamma_0(N)\backslash \operatorname{GSp}(4, \mathbb{Q})/Q(\mathbb{Q})).$$

Let

$$\omega_1(q) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{for } q = \begin{bmatrix} * & * & * & * \\ & a & b & * \\ & c & d & * \\ & & & * \end{bmatrix} \in Q(\mathbb{Q}),$$

and let ι_1 be an embedding map

$$\iota_1\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \begin{bmatrix} 1 & * & * & * \\ & a & b & * \\ & c & d & * \\ & & & 1 \end{bmatrix}$$

from $\operatorname{SL}(2, \mathbb{Q})$ to $Q(\mathbb{Q})$. For $g \in \operatorname{GSp}(4, \mathbb{Q})$, we define $\Gamma_g := \omega_1(g^{-1}\Gamma_0(\mathrm{N})g \cap Q(\mathbb{Q}))$.

3. A brief overview of the main results

We recall cusps in the degree one case. Let Γ be a congruence subgroup of $\operatorname{SL}(2, \mathbb{Z})$ which acts on the complex upper half plane $\mathbb{H}$ by the usual action. In order to compactify $\Gamma\backslash\mathbb{H}$ we adjoin $\mathbb{Q}\cup\{\infty\}$ to $\mathbb{H}$ to define the extended plane $\mathbb{H}^* = \mathbb{H}\cup\mathbb{Q}\cup\{\infty\}$ and take the quotient $X(\Gamma) = \Gamma\backslash\mathbb{H}^*$. Then a cusp of $X(\Gamma)$ is a Γ-equivalence class of points in $\mathbb{Q}\cup\{\infty\}$. As $\operatorname{SL}(2, \mathbb{Z})$ acts transitively on $\mathbb{Q}\cup\{\infty\}$ there is just one cusp of the modular curve $X(1) = \operatorname{SL}(2, \mathbb{Z})\backslash\mathbb{H}^*$. It is well known that cusps of $X(\Gamma_0^2(N))$ correspond to the double coset decompositions of $\Gamma_0^2(N)\backslash \operatorname{SL}(2, \mathbb{Z})/\Gamma_\infty^2(\mathbb{Z})$, for example see [Diamond and Shurman 2005, Proposition 3.8.5] or [Miyake 1989, §4.2].

The theory of Satake compactification is explained in [Satake 1957/58a]. A quick review can be found in [Poor and Yuen 2013, Section 3]. In fact similar to the degree one case, the one-dimensional cusps for the Siegel congruence subgroup $\Gamma_0(N)$, in the degree two case, correspond to the double coset decompositions $\Gamma_0(N)\backslash \operatorname{Sp}(4, \mathbb{Z})/\Gamma_\infty(\mathbb{Z})$ and also equivalently to $\Gamma_0(N)\backslash \operatorname{GSp}(4, \mathbb{Q})/Q(\mathbb{Q})$. Similarly the zero-dimensional cusps correspond to the double coset decompositions $\Gamma_0(N)\backslash \operatorname{GSp}(4, \mathbb{Q})/P(\mathbb{Q})$. It turns out that for even weights $k > 4$, the codimension of cusp forms can be obtained by using the Satake's theorem; see [Satake 1957/58b] if the structure of zero-dimensional cusps and one-dimensional cusps are known.

We prove the following result concerning one-dimensional cusps in the case when $N = p^n$ for some prime p and $n \geq 1$. In fact, the one-dimensional cusps for $\Gamma_0(p^n)$ are inverses of the representatives listed below.

Theorem 3.1. *Assume $n \geq 1$. A complete and minimal system of representatives for the double cosets $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p^n)$ is given by*

$$1, \quad s_1 s_2, \quad g_1(p, \gamma, r) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ \gamma p^r & & & 1 \end{bmatrix}$$

$$g_2(p, s) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ & p^s & & 1 \end{bmatrix}, \quad g_3(p, \delta, r, s) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ \delta p^r & p^s & & 1 \end{bmatrix}$$

where $1 \leq s, r \leq n-1$, $s < r < 2s$ and where γ, δ run through elements in $(\mathbb{Z}/p^{f_1}\mathbb{Z})^\times$ and $(\mathbb{Z}/p^{f_2}\mathbb{Z})^\times$, respectively, with $f_1 = \min(r, n-r)$ and $f_2 = \min(2s - r, n - r)$. The total number of representatives given above is

$$C_1(p^n) = \begin{cases} \frac{p^{n/2+1} + p^{n/2} - 2}{p-1} & \text{if } n \text{ is even,} \\ \frac{2(p^{n+1/2} - 1)}{p-1} & \text{if } n \text{ is odd.} \end{cases} \tag{3-1}$$

Some remarks.

(i) We note that alternatively one can get the following system of complete and minimal representatives for the double cosets $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p^n)$.

$$g_1(\gamma, x) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ x\gamma & & & 1 \end{bmatrix}, \quad 1 \leq \gamma \leq N, \ \gamma \mid N,$$

$$g_3(\gamma, \delta, y) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ \delta & & 1 & \\ y\gamma & \delta & & 1 \end{bmatrix}, \quad 1 < \delta < \gamma \leq N, \ \gamma \mid N, \ \delta \mid N, \ \delta \mid \gamma, \ \gamma \mid \delta^2;$$

where $N = p^n$ and for fixed γ and δ we set

$$M = \gcd\Big(\gamma, \frac{N}{\gamma}\Big), \quad L = \gcd\Big(\frac{\delta^2}{\gamma}, \frac{N}{\gamma}\Big),$$

and x and y vary through all the elements of $(\mathbb{Z}/M\mathbb{Z})^\times$ and $(\mathbb{Z}/L\mathbb{Z})^\times$, respectively. Here we interpret $(\mathbb{Z}/\mathbb{Z})^\times$ as an empty set. Clearly $g_1(N, x)$ is equivalent to the representative 1 in $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p^n)$ and one can show that $g_1(1, x)$ is equivalent to the representative $s_1 s_2$ (see Lemma 5.7). One can also easily show that $g_3(N, p^s, 1)$ is equivalent to the representative $g_2(p, s)$.

(ii) One can write yet another formulation for a complete and minimal system of representatives for the double cosets $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p^n)$ as follows:

$$g_0(\gamma, \delta, y) := \begin{bmatrix} 1 & & & \\ & 1 & & \\ \delta & & 1 & \\ y\gamma & \delta & & 1 \end{bmatrix}, \quad 1 \le \delta \le \gamma \le N,\ \gamma \mid N,\ \delta \mid N,\ \delta \mid \gamma,\ \gamma \mid \delta^2,$$

with y, γ, δ and N as in the first remark. This is clear on observing that the definition of $g_0(\gamma, \delta, y)$ is different from $g_3(\gamma, \delta, y)$ only when $\delta = \gamma$ and in that case the set of representatives $g_0(\gamma, \gamma, y)$ is equivalent to $g_1(\gamma, x)$.

The above result can be extended by using the strong approximation theorem and the Chinese remainder theorem to arbitrary N. We have the following lemma.

Lemma 3.2. *Assume $N = \prod_{i=1}^m p_i^{n_i}$. Then, the number of inequivalent representatives for the double cosets $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(\mathrm{N})$ is given by $C_1(N) = \prod_{i=1}^m C_1(p_i^{n_i})$.*

We have the following corollary of Theorem 3.1 based on Lemma 3.2.

Corollary 3.3. *Assume $N = \prod_{i=1}^m p_i^{n_i}$. A complete and minimal system of representatives for the double cosets $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/ \Gamma_0(N)$ is given by*

$$g_1(\gamma, x) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ x\gamma & & & 1 \end{bmatrix}, \quad 1 \le \gamma \le N,\ \gamma \mid N,$$

$$g_3(\gamma, \delta, y) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ \delta & & 1 & \\ y\gamma & \delta & & 1 \end{bmatrix}, \quad 1 < \delta < \gamma \le N,\ \gamma \mid N,\ \delta \mid N,\ \delta \mid \gamma,\ \gamma \mid \delta^2,$$

where for fixed γ and δ we have

$$x = M + \zeta \prod_{p_i \nmid M,\ p_i \mid N} p_i^{n_i}, \qquad y = L + \theta \prod_{p_i \nmid L,\ p_i \mid N} p_i^{n_i},$$

with $M = \gcd(\gamma, N/\gamma)$, $L = \gcd(\delta^2/\gamma, N/\gamma)$, ζ and θ varies through all the elements of $(\mathbb{Z}/M\mathbb{Z})^\times$ and $(\mathbb{Z}/L\mathbb{Z})^\times$, respectively. Here we interpret $(\mathbb{Z}/\mathbb{Z})^\times$ as an empty set.

Essentially the representatives listed above are obtained by appropriately lifting the representatives of $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/ \Gamma_0(p_i^{n_i})$ for each prime factor p_i of N. The last statement will be made explicit in the proof of the corollary. We also note that x and y are defined in such a way that $\gcd(x, N) = 1$ and $\gcd(y, N) = 1$.

We remark that the one-dimensional cusps for $\Gamma_0(N)$ are given by the inverses of the representatives listed above in Corollary 3.3.

Corollary 3.4. (1) *Let f_1 be an elliptic cusp form of even weight k with $k \geq 6$ and level N. Let $\jmath(g)$ be a one-dimensional cusp for $\Gamma_0(N)$ of the form $g_1(\gamma, x)^{-1}$. Then*

$$E_g(Z) = \sum_{\jmath(\xi) \in (\jmath(g) Q(\mathbb{Q}) \jmath(g^{-1}) \cap \Gamma_0(N)) \backslash \Gamma_0(N)} f_1(g^{-1}\xi \langle Z \rangle^*) \det(j(g^{-1}\xi, Z))^{-k}$$

defines a Klingen–Eisenstein series of level N with respect to the Siegel congruence subgroup $\Gamma_0(N)$.

(2) *Let $\jmath(h)$ be a one-dimensional cusp for $\Gamma_0(N)$ of the form $g_3(\gamma, \delta, y)^{-1}$. Let $f_2 \in S_k(\Gamma_{\jmath(h)})$ with even weight k such that $k \geq 6$. Then*

$$E_h(Z) = \sum_{\jmath(\xi) \in (\jmath(h) Q(\mathbb{Q}) \jmath(h^{-1}) \cap \Gamma_0(N)) \backslash \Gamma_0(N)} f_2(h^{-1}\xi \langle Z \rangle^*) \det(j(h^{-1}\xi, Z))^{-k}$$

defines a Klingen–Eisenstein series of level N with respect to the Siegel congruence subgroup $\Gamma_0(N)$.

As $\jmath(g)$ and $\jmath(h)$ run through all one-dimensional cusps of the form $g_1(\gamma, x)^{-1}$ and $g_3(\gamma, \delta, y)^{-1}$ respectively, and for some fixed g and h, as f_1 and f_2 run through a basis of $S_k(\Gamma_0(N))$ and $S_k(\Gamma_{\jmath(h)})$ respectively, the Klingen–Eisenstein series thus obtained are linearly independent.

The number of zero-dimensional cusps $C_0(p^n)$ for odd prime p was calculated by Markus Klein in his thesis [2004, Korollar 2.28]:

$$C_0(p^n) = 2n + 1 + 2\left(\sum_{j=1}^{n-1} \phi(p^{\min(j,n-j)}) + \sum_{j=1}^{n-2} \sum_{i=j+1}^{n-1} \phi(p^{\min(j,n-i)})\right). \tag{3-2}$$

It is the same as

$$C_0(p^n) = \begin{cases} 3 & \text{if } n = 1, \\ 2p + 3 & \text{if } n = 2, \\ -2n - 1 + 2p^{n/2} + 8\frac{p^{n/2}-1}{p-1} & \text{if } n \geq 4 \text{ is even}, \\ -2n - 1 + 6p^{(n-1)/2} + 8\frac{p^{(n-1)/2}-1}{p-1} & \text{if } n \geq 3 \text{ is odd}. \end{cases} \tag{3-3}$$

The above formula remains valid if $p = 2$ and $n = 1$. The above result also remains true for $p = 2$ and $n = 2$ as calculated by Tsushima (cf. [Tsushima 2003]). Hence, assume $8 \nmid N$ and if $N = \prod_{i=1}^{m} p_i^{n_i}$ then following an argument similar to the one given in the proof of Lemma 3.2 we obtain

$$C_0(N) = \prod_{i=1}^{m} C_0(p_i^{n_i}). \tag{3-4}$$

Finally, by using Satake's [1957/58b] theorem and the formula for $C_0(N)$ and $C_1(N)$ described above we obtain the following result:

Theorem 3.5. *Let $N \geq 1$, $8 \nmid N$ and $k \geq 6$, even. Then*

$$\dim M_k(\Gamma_0(N)) - \dim S_k(\Gamma_0(N)) = C_0(N) + \Bigl(\sum_{\gamma \mid N} \phi(\gcd(\gamma, N/\gamma))\Bigr) \dim S_k(\Gamma_0^2(N)) + \sum_{\substack{1<\delta<\gamma,\ \gamma \mid N,\\ \delta\mid\gamma,\ \gamma\mid\delta^2}} {\sum}' \dim S_k(\Gamma_g) \tag{3-5}$$

where $C_0(N)$ is given by (3-4) if $N > 1$, $C_0(1) = 1$, ϕ is Euler's totient function, and for a fixed γ and δ, the summation $\sum'$ is carried out such that g runs through every one-dimensional cusp of the form $g_3(\gamma, \delta, y)$, with y taking all possible values as in Corollary 3.3.

Some remarks.

(i) We note that Markus Klein did not consider the case $4 \mid N$ for calculating the number of zero-dimensional cusps in his thesis. Tsushima provided the result for $N = 4$. Since we refer to their results for the number of zero-dimensional cusps we have this restriction in our theorem. We hope to return to this case in the future.

(ii) The above result in the special case of square-free N reduces to the dimension formula given in [Böcherer and Ibukiyama 2012] for even $k \geq 6$. They also treat the case $k = 4$ for square-free N.

In Section 4 we briefly review Satake compactification and cusps. Thereafter in Section 5 we give proofs of the main results. We remark that the proof of Theorem 3.1 is entirely algorithmic and essentially uses elementary number theory to establish the result.

4. Cusps of $\Gamma_0(N)$

We recall a few basic facts related to the Satake compactification $S(\Gamma\backslash\mathbb{H}_2)$ of $\Gamma\backslash\mathbb{H}_2$ (see [Satake 1957/58a; 1957/58b; Böcherer and Ibukiyama 2012; Poor and Yuen 2013]). Here Γ is a congruence subgroup of $\mathrm{Sp}(4, \mathbb{Z})$. We will be interested in $S(N) := S(\Gamma_0(N)\backslash\mathbb{H}_2)$. By $\mathrm{Bd}(N)$ we denote the boundary of $S(N)$. The one-dimensional components of $\mathrm{Bd}(N)$ are modular curves and are called the one-dimensional cusps. The one-dimensional cusps intersect on the zero-dimensional cusps.

We define $M_k(\mathrm{Bd}(N))$ to be the space of modular forms on $\mathrm{Bd}(N)$ which consists of modular forms of weight k on the one-dimensional boundary components such that they are compatible on each intersection point. In the following we make the above description more explicit. Let $\mathrm{GSp}(4, \mathbb{Q}) = \bigsqcup_{i=1}^{l} \Gamma_0(\mathrm{N}) g_i Q(\mathbb{Q})$.

Then the one-dimensional cusps bijectively correspond to $\{g_i\}$. Let $\Gamma_i = \omega_1(g_i^{-1}\Gamma_0(\mathrm{N})g_i \cap Q(\mathbb{Q}))$. In this situation the one-dimensional cusp g_i can be associated to the modular curve $\Gamma_i\backslash\mathbb{H}_1$. The zero-dimensional cusps of $\Gamma_i\backslash\mathbb{H}_1$ correspond to the representatives h_j of $\Gamma_i \backslash \mathrm{SL}(2, \mathbb{Z}) / \Gamma_\infty^2(\mathbb{Z})$. In fact, h_j can be identified with the zero-dimensional cusp of $S(N)$ that corresponds to $\Gamma_0(N)g_i\iota_1(h_j)P(\mathbb{Q})$. If $\Gamma_0(N)g_i\iota_1(h_j)P(\mathbb{Q}) = \Gamma_0(N)g_r\iota_1(h_j)P(\mathbb{Q})$ for two inequivalent one-dimensional cusps g_i and g_r then it means that these two one-dimensional cusps intersect at a zero-dimensional cusp. Next we define a map Φ from $M_k(\Gamma_0(N))$ to $M_k(\Gamma_i)$ by $(\Phi(F))(z) = \lim_{\lambda\to\infty} F\left(\left[\begin{smallmatrix} z & \\ & i\lambda \end{smallmatrix}\right]\right)$ for $z \in \mathbb{H}_1$. Then we define $\tilde{\Phi} : M_k(\Gamma_0(N)) \to M_k(\mathrm{Bd}(N))$ by $F \to (\Phi(F|_k\jmath(g_i)))_{1\le i\le l}$. Here $|_k$ denotes the usual slash operator defined as $F|_k\jmath(g) = \det(CZ+D)^{-k}F(\jmath(g)\langle Z\rangle)$ for $\jmath(g) = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ and $\jmath(g)\langle Z\rangle = (AZ + B)(CZ + D)^{-1}$ with $g \in \mathrm{Sp}(4, \mathbb{R})$. Any element $(f_i)_{1\le i\le l}$ in the image of $\tilde{\Phi}$ satisfies the condition that $f_i|_k h_1 = f_j|_k h_2$ whenever $\Gamma_0(N)g_i\iota_1(h_1)P(\mathbb{Q}) = \Gamma_0(N)g_j\iota_1(h_2)P(\mathbb{Q})$; where $h_1, h_2 \in \mathrm{SL}(2, \mathbb{Q})$ and $1 \le i, j \le l$. It essentially means that f_i and f_j, which are modular forms on the one-dimensional cusps g_i and g_j respectively, are compatible on the intersection points of these cusps.

5. Proofs

Proof of Theorem 3.5.

Proof. By Satake's theorem (see [Satake 1957/58b]) it follows that the codimension of the space of cusp forms is $\dim M_k(\mathrm{Bd}(N))$. We recall that by definition $f \in M_k(\mathrm{Bd}(N))$ means: f is a modular form of weight k on the boundary components of $S(N)$ such that f takes the same value on each intersection point of the boundary components. If $f \in S_k(\Gamma_i)$ on a boundary component $\Gamma_i\backslash\mathbb{H}_1$ corresponding to a one-dimensional cusp, say g_i, then f vanishes at every cusp of g_i and in particular f takes the same value zero at every intersection point of the boundary components. Hence $f \in M_k(\mathrm{Bd}(N))$. We note that for any representatives g_1 of the form $g_1(\gamma, x)$ with $g_1(\gamma, x)$ defined as in Corollary 3.3, we have $\omega_1(g_1^{-1}\Gamma_0(\mathrm{N})g_1 \cap Q(\mathbb{Q})) = \Gamma_0^2(N)$ and similarly for any representatives g_3 of the form $g_3(\gamma, \delta, y)$ a simple calculation shows that $\Gamma_{g_3} = \omega_1(g_3^{-1}\Gamma_0(\mathrm{N})g_3 \cap Q(\mathbb{Q})) \subset \Gamma_0^2(\delta)$. It follows that each one-dimensional cusp of the form $g_1(\gamma, x)$ contributes $\dim S_k(\Gamma_0^2(N))$ linearly independent cusp forms and this accounts for the second term in the formula (3-5). For a fixed δ and γ such that $1 < \delta < \gamma$, $\gamma \mid N$, $\delta \mid \gamma$, $\gamma \mid \delta^2$ and for a fixed y, the one-dimensional cusp g of the form $g_3(\gamma, \delta, y)$ contributes $\dim S_k(\Gamma_g)$ cusp forms. These contributions account for the last term in the summation formula (3-5). We remark that the last two terms in the summation formula (3-5) count the Klingen–Eisenstein series associated to each one-dimensional cusp as defined in

Corollary 3.4. Finally, since $k > 4$ and even there exists a basis of Eisenstein series that is supported at a single cusp. The total number of zero-dimensional cusps $C_0(N)$ accounts for all such cases and these are in fact Siegel–Eisenstein series. □

In the following we determine a complete and minimal system of representatives for the double cosets $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p^n)$ and prove Theorem 3.1. For that we begin with first stating and proving several lemmas. In the following we write $g \sim h \Longleftrightarrow Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})h\Gamma_0(p^n)$, for $g, h \in \mathrm{GSp}(4, \mathbb{Q})$.

Lemma 5.1. *Let p be a prime number and let x_1, x_2 be integers such that x_1, x_2 and p are pairwise coprime. Further, assume y_1, y_2 to be integers such that y_1, y_2 and p are pairwise coprime with $\gcd(x_1, y_1) = 1$. Let $x = x_1x_2^{-1}p^{-r}$ and $y = y_1y_2^{-1}p^{-s}$ with $r, s \geq 0$, $x_2 \neq 0$, $y_2 \neq 0$. Let $n \geq 1$. Let*

$$g = s_1s_2 \begin{bmatrix} 1 & & y & \\ & 1 & x & y \\ & & 1 & \\ & & & 1 \end{bmatrix}.$$

Then we have the following results.

(1) *If $s > r$, then there exist integers η_1 and η_2 which are coprime to p such that*

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ \eta_1 p^s & & 1 & \\ \eta_2 p^{-r+2s} & \eta_1 p^s & & 1 \end{bmatrix} \Gamma_0(p^n).$$

(2) *If $s \leq r < n$, then there exists a nonzero integer x_3 coprime to p such that*

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ x_3 p^r & & & 1 \end{bmatrix} \Gamma_0(p^n),$$

and if $s \leq r$ and $r \geq n$, then $Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})1\Gamma_0(p^n)$.

Proof. We have

$$s_1s_2 \begin{bmatrix} 1 & & y & \\ & 1 & x & y \\ & & 1 & \\ & & & 1 \end{bmatrix} \sim \begin{bmatrix} -x & & -y & -1 \\ & & -1 & \\ & 1 & y^2/x & y/x \\ & & & -1/x \end{bmatrix} s_1s_2 \begin{bmatrix} 1 & & y & \\ & 1 & x & y \\ & & 1 & \\ & & & 1 \end{bmatrix} s_1$$

$$= \begin{bmatrix} 1 & & & \\ & & -1 & \\ -y/x & 1 & & \\ 1/x & & y/x & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & & & \\ & & 1 & \\ -y/x & & & 1 \\ 1/x & -y/x & & 1 \end{bmatrix} s_2.$$

Now we prove the first part of the lemma.

Case 1: $s > r$ Assume $s > r$. Let $\gcd(y_2, x_2) = \tau$. Let l_1, l_2 be integers such that

$$l_1 x_2 y_1 + l_2 p^{s-r} x_1 y_2 = \tau.$$

Let

$$d_1 = \frac{l_1^2 x_2 y_2 p^s}{\tau}, \quad d_2 = \frac{l_1 l_2 x_2 y_2 p^s}{\tau}.$$

It follows that

$$d_1 x_2 y_1 + d_2 x_1 y_2 p^{s-r} = l_1 p^s x_2 y_2.$$

Then we have

$$\begin{bmatrix} 1 & & & \\ & & -1 & \\ -\frac{p^r x_2 y_1}{p^s x_1 y_2} & 1 & & \\ \frac{p^r x_2}{x_1} & & \frac{p^r x_2 y_1}{p^s x_1 y_2} & 1 \end{bmatrix}$$

$$\sim \begin{bmatrix} \frac{\tau}{p^{-r+s} x_1 y_2} & & l_1 & \\ & & \frac{p^{-r+s} x_1 y_2}{\tau} & \\ & -\frac{\tau}{p^{-r+s} x_1 y_2} & -d_1 & -l_1 \\ & & & \frac{p^{-r+s} x_1 y_2}{\tau} \end{bmatrix} \begin{bmatrix} 1 & & & \\ & & -1 & \\ -\frac{p^r x_2 y_1}{p^s x_1 y_2} & 1 & & \\ \frac{p^r x_2}{x_1} & & \frac{p^r x_2 y_1}{p^s x_1 y_2} & 1 \end{bmatrix}$$

$$\sim \begin{bmatrix} -\frac{l_1 p^r x_2 y_1}{p^s x_1 y_2} + \frac{\tau}{p^{-r+s} x_1 y_2} & -l_1 & & \\ -\frac{x_2 y_1}{\tau} & -\frac{p^{-r+s} x_1 y_2}{\tau} & & \\ -\frac{l_1 p^r x_2}{x_1} + \frac{d_1 p^r x_2 y_1}{p^s x_1 y_2} & d_1 & \frac{l_1 p^r x_2 y_1}{p^s x_1 y_2} - \frac{\tau}{p^{-r+s} x_1 y_2} & -l_1 \\ \frac{p^s x_2 y_2}{\tau} & & -\frac{x_2 y_1}{\tau} & \frac{p^{-r+s} x_1 y_2}{\tau} \end{bmatrix}$$

$$\sim \begin{bmatrix} l_2 & -l_1 & & \\ -\frac{x y_1}{\tau} & -\frac{p^{-r+s} x_1 y_2}{\tau} & & \\ -d_2 & d_1 & -l_2 & -l_1 \\ \frac{p^s x_2 y_2}{\tau} & & -\frac{x_2 y_1}{\tau} & \frac{p^{-r+s} x_1 y_2}{\tau} \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & & & \\ & 1 & & \\ -\frac{l_1 p^s x_2 y_2}{\tau} & & 1 & \\ \frac{p^{-r+2s} x_1 x_2 y_2^2}{\tau^2} & -\frac{l_1 p^s x_2 y_2}{\tau} & & 1 \end{bmatrix}$$

$$\begin{bmatrix} l_2 & -l_1 & & \\ -\frac{x_2 y_1}{\tau} & -\frac{p^{-r+s} x_1 y_2}{\tau} & & \\ & & -l_2 & -l_1 \\ -\frac{l_1 p^s x_2^2 y_1 y_2}{\tau^2} - \frac{l_2 p^{-r+2s} x_1 x_2 y_2^2}{\tau^2} + \frac{p^s x_2 y_2}{\tau} & & -\frac{x_2 y_1}{\tau} & \frac{p^{-r+s} x_1 y_2}{\tau} \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & & & \\ & 1 & & \\ -\frac{l_1 p^s x_2 y_2}{\tau} & & 1 & \\ \frac{p^{-r+2s} x_1 x_2 y_2^2}{\tau^2} & -\frac{l_1 p^s x_2 y_2}{\tau} & & 1 \end{bmatrix} \begin{bmatrix} l_2 & -l_1 & & \\ -\frac{x_2 y_1}{\tau} & -\frac{p^{-r+s} x_1 y_2}{\tau} & & \\ & & -l_2 & -l_1 \\ & & -\frac{x_2 y_1}{\tau} & \frac{p^{-r+s} x_1 y_2}{\tau} \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & & & \\ & 1 & & \\ -\frac{l_1 p^s x_2 y_2}{\tau} & & 1 & \\ \frac{p^{-r+2s} x_1 x_2 y_2^2}{\tau^2} & -\frac{l_1 p^s x_2 y_2}{\tau} & & 1 \end{bmatrix}.$$

This completes the proof of the first part of the lemma with $\eta_1 = -l_1 x_2 y_2/\tau$ and $\eta_2 = x_1 x_2 y_2^2/\tau^2$.

Now we prove the second part of the lemma.

Case 2: $s \le r$ Assume $s \le r$ and $\gcd(x_1, y_1) = 1$. Let $\gcd(y_2, x_2) = \tau$. Let l_1 and l_2 be integers such that

$$l_1 x_2 y_1 p^{r-s} + l_2 x_1 y_2 = \tau.$$

Let d_1 and d_2 be integers such that

$$d_1 y_1 + d_2 x_1 p^s = -l_1 p^s.$$

Let c_1 and c_2 be integers such that

$$c_1 \frac{x_1 y_2}{\tau} + c_2 p^{\max(0, n-s)} = \frac{-d_1 y_2}{p^s}.$$

Let

$$\beta = \frac{x_2}{\tau}(d_2 \tau - c_1 y_1).$$

It is easy to see that $\beta \in \mathbb{Z}$, as $\tau = \gcd(y_2, x_2)$. Further, if $r < n$ then we make the following choices. Let l_3 and l_4 be integers such that

$$l_3 \frac{x_2 y_2}{\tau} + l_4 p^{n-r} = \beta.$$

Let x_3 and x_4 be integers such that

$$x_3 \left(l_3 p^{r-s} \frac{x_2 y_1}{\tau} \right) - x_4 p^{n-r} = -\frac{x_2 y_2}{\tau}.$$

Otherwise, if $r \ge n$ then let

$$l_3 = l_4 = x_3 = x_4 = 0.$$

With the above choices in place, we obtain

$$
\begin{bmatrix} \frac{\tau}{x_1y_2} & & l_1 & \\ & & \frac{x_1y_2}{\tau} & \\ & -\frac{\tau}{x_1y_2} & d_1y_2 & -l_1 \\ & & & \frac{x_1y_2}{\tau} \end{bmatrix}
\begin{bmatrix} 1 & & & \\ & & -1 & \\ -\frac{p^rx_2y_1}{p^sx_1y_2} & 1 & & \\ \frac{p^rx_2}{x_1} & & \frac{p^rx_2y_1}{p^sx_1y_2} & 1 \end{bmatrix}
$$

$$
\sim \begin{bmatrix} -\frac{l_1p^rx_2y_1}{p^sx_1y_2}+\frac{\tau}{x_1y_2} & -l_1 & & \\ -\frac{p^rx_2y_1}{p^s\tau} & -\frac{x_1y_2}{\tau} & & \\ -\frac{l_1p^rx_2}{x_1}-\frac{d_1p^rx_2y_1}{p^sx_1} & -d_1y_2 & \frac{l_1p^rx_2y_1}{p^sx_1y_2}-\frac{\tau}{x_1y_2} & -l_1 \\ \frac{p^rx_2y_2}{\tau} & & -\frac{p^rx_2y_1}{p^s\tau} & \frac{x_1y_2}{\tau} \end{bmatrix}
$$

$$
\sim \begin{bmatrix} l_2 & -l_1 & & \\ -\frac{p^rx_2y_1}{p^s\tau} & -\frac{x_1y_2}{\tau} & & \\ d_2p^rx_2 & -d_1y_2 & -l_2 & -l_1 \\ \frac{p^rx_2y_2}{\tau} & & -\frac{p^rx_2y_1}{p^s\tau} & \frac{x_1y_2}{\tau} \end{bmatrix}
$$

$$
\sim \begin{bmatrix} 1 & l_3 & & \\ & 1 & & \\ & c_1p^s & 1 & -l_3 \\ & & & 1 \end{bmatrix}
\begin{bmatrix} l_2 & -l_1 & & \\ -\frac{p^rx_2y_1}{p^s\tau} & -\frac{x_1y_2}{\tau} & & \\ d_2p^rx_2 & -d_1y_2 & -l_2 & -l_1 \\ \frac{p^rx_2y_2}{\tau} & & -\frac{p^rx_2y_1}{p^s\tau} & \frac{x_1y_2}{\tau} \end{bmatrix}
$$

$$
\sim \begin{bmatrix} -\frac{l_3p^rx_2y_1}{p^s\tau}+l_2 & -\frac{l_3x_1y_2}{\tau}-l_1 & & \\ -\frac{p^rx_2y_1}{p^s\tau} & -\frac{x_1y_2}{\tau} & & \\ d_2p^rx_2-\frac{c_1p^rx_2y_1}{\tau}-\frac{l_3p^rx_2y_2}{\tau} & -\frac{c_1p^sx_1y_2}{\tau}-d_1y_2 & \frac{l_3p^rx_2y_1}{p^s\tau}-l_2 & -\frac{l_3x_1y_2}{\tau}-l_1 \\ \frac{p^rx_2y_2}{\tau} & & -\frac{p^rx_2y_1}{p^s\tau} & \frac{x_1y_2}{\tau} \end{bmatrix}
$$

$$
\sim \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ p^rx_3 & & & 1 \end{bmatrix}
$$

$$
\begin{bmatrix} -\frac{l_3p^rx_2y_1}{p^s\tau}+l_2 & -\frac{l_3x_1y_2}{\tau}-l_1 & & \\ -\frac{p^rx_2y_1}{p^s\tau} & -\frac{x_1y_2}{\tau} & & \\ d_2p^rx_2-\frac{c_1p^rx_2y_1}{\tau}-\frac{l_3p^rx_2y_2}{\tau} & -\frac{c_1p^sx_1y_2}{\tau}-d_1y_2 & \frac{l_3p^rx_2y_1}{p^s\tau}-l_2 & -\frac{l_3x_1y_2}{\tau}-l_1 \\ \frac{l_3(p^r)^2x_2x_3y_1}{p^s\tau}-l_2p^rx_3+\frac{p^rx_2y_2}{\tau} & \frac{l_3p^rx_1x_3y_2}{\tau}+l_1p^rx_3 & -\frac{p^rx_2y_1}{p^s\tau} & \frac{x_1y_2}{\tau} \end{bmatrix}
$$

$$
\sim \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ p^rx_3 & & & 1 \end{bmatrix},
$$

and the second part of the lemma follows. □

Lemma 5.2. *Assume p to be a prime number and n be a positive integer. Let*

$$x=\frac{x_1}{p^r x_2},\qquad y=\frac{y_1}{p^s y_2}\quad \textit{and}\quad z=\frac{z_1}{p^t z_2},$$

where r, s, t are nonnegative integers, x_1, x_2, y_2, z_2 are nonzero integers and y_1, z_1 are integers. Let any two nonzero elements from the set $\{x_1, y_1, z_1, x_2, y_2, z_2, p\}$ be mutually coprime except, possibly, when both belong to $\{x_2, y_2, z_2\}$. Let

$$g=s_1s_2s_1\begin{bmatrix}1&x&&\\&1&&\\&&1&-x\\&&&1\end{bmatrix}\begin{bmatrix}1&&y&z\\&1&&y\\&&1&\\&&&1\end{bmatrix}.$$

Then there exist $x', y'\in\mathbb{Q}$ such that

$$Q(\mathbb{Q})g\Gamma_0(p^n)=Q(\mathbb{Q})s_1s_2\begin{bmatrix}1&&y'&\\&1&x'&y'\\&&1&\\&&&1\end{bmatrix}\Gamma_0(p^n).$$

Proof. We have

$$\begin{aligned}
g&\sim s_1s_2s_1\begin{bmatrix}1&x&&\\&1&&\\&&1&-x\\&&&1\end{bmatrix}\begin{bmatrix}1&&y&z\\&1&&y\\&&1&\\&&&1\end{bmatrix}s_1\\
&=\begin{bmatrix}1&&&\\&1&&\\-x&&1&\\&-x&&1\end{bmatrix}s_1s_2\begin{bmatrix}1&&y&\\&1&z&y\\&&1&\\&&&1\end{bmatrix}\\
&\sim\begin{bmatrix}x&&1&\\&x&&1\\&&\frac{1}{x}&\\&&&\frac{1}{x}\end{bmatrix}\begin{bmatrix}1&&&\\&1&&\\-x&&1&\\&-x&&1\end{bmatrix}s_1s_2\begin{bmatrix}1&&y&\\&1&z&y\\&&1&\\&&&1\end{bmatrix}(s_1)^{-1}\\
&=s_2s_1s_2s_1s_2\begin{bmatrix}1&x^{-1}&&\\&1&&\\&&1&-x^{-1}\\&&&1\end{bmatrix}\begin{bmatrix}1&&y&\\&1&z&y\\&&1&\\&&&1\end{bmatrix}.
\end{aligned}\tag{5-1}$$

Case 1: $y \neq 0, z \neq 0$ Let us first consider the case when $y \neq 0$ and $z \neq 0$. Let $\alpha = \gcd(x_2 y_2, z_2)$. Further, if $s > t - r$, then let

$$
\begin{cases}
-d_1 p^{n+t} + d_2 x_1 & = 1 \quad \text{with } d_1, d_2 \in \mathbb{Z}, \\
d_3 & = d_1 x_2 y_2 z_2 p^{n+s+r+t}, \\
c_1 & = (x_2 y_2 z_1 p^{r+s-t} + x_1 y_1 z_2)\alpha^{-1}, \\
\tau & = \gcd(d_3, c_1), \\
y_4 & = \alpha \tau x_1^{-1} z_2^{-1} p^{-s}, \\
-d_4 \tau p^n + d_5 x_1 & = 1 \quad \text{with } d_4, d_5 \in \mathbb{Z}, \\
d & = d_4 x_2 y_2 z_2 \alpha^{-1} p^{n+s+r}.
\end{cases}
\tag{5-2}
$$

Otherwise, if $s \leq t - r$ then let

$$
\begin{cases}
-d_1 p^{n+s+r} + d_2 x_1 & = 1 \quad \text{with } d_1, d_2 \in \mathbb{Z}, \\
d_3 & = d_1 x_2 y_2 z_2 p^{n+s+r+t}, \\
c_1 & = (x_2 y_2 z_1 + x_1 y_1 z_2 p^{-r-s+t})\alpha^{-1}, \\
\tau & = \gcd(d_3, c_1), \\
y_4 & = \alpha \tau x_1^{-1} z_2^{-1} p^{r-t}, \\
-d_4 \tau p^n + d_5 x_1 & = 1 \quad \text{with } d_4, d_5 \in \mathbb{Z}, \\
d & = d_4 x_2 y_2 z_2 \alpha^{-1} p^{n+t}.
\end{cases}
\tag{5-3}
$$

Now, if $\tau > 1$ then we replace c_1 by $c_1 \tau^{-1}$. Then, if needed on appropriately adjusting d_4 and d_5, we pick integers a_1 and b such that

$$a_1 d - b c_1 = 1.$$

Next, we set

$$
\begin{aligned}
a &= a_1 + b z_1 z_2^{-1} p^{-t}, \\
c &= c_1 + d z_1 z_2^{-1} p^{-t}, \\
x_4 &= \frac{c p^r x_2 y_4}{x_1} + \frac{p^r x_2 y_1}{p^s x_1} + \frac{d y_1 y_4}{p^s y_2}.
\end{aligned}
$$

Then

$$
\begin{aligned}
g &\sim s_2s_1s_2s_1s_2 \begin{bmatrix} 1 & x^{-1} & & \\ & 1 & & \\ & & 1 & -x^{-1} \\ & & & 1 \end{bmatrix} \begin{bmatrix} 1 & & y & \\ & 1 & z & y \\ & & 1 & \\ & & & 1 \end{bmatrix} \qquad \text{(from (5-1))} \\
&\sim \begin{bmatrix} & & & 1 \\ & & 1 & -\frac{p^r x_2}{x_1} \\ & -1 & -\frac{z_1}{p^t z_2} & -\frac{y_1}{p^s y_2} \\ -1 & -\frac{p^r x_2}{x_1} & -\frac{p^r x_2 z_1}{p^t x_1 z_2} - \frac{y_1}{p^s y_2} & -\frac{p^r x_2 y_1}{p^s x_1 y_2} \end{bmatrix} \\
&\sim \begin{bmatrix} 1 & & & \\ & a & b & \\ & c & d & \\ & & & 1 \end{bmatrix} \begin{bmatrix} & & & 1 \\ & & 1 & -\frac{p^r x_2}{x_1} \\ & -1 & -\frac{z_1}{p^t z_2} & -\frac{y_1}{p^s y_2} \\ -1 & -\frac{p^r x_2}{x_1} & -\frac{p^r x_2 z_1}{p^t x_1 z_2} - \frac{y_1}{p^s y_2} & -\frac{p^r x_2 y_1}{p^s x_1 y_2} \end{bmatrix} \\
&\sim \begin{bmatrix} & & & 1 \\ & -b & a - \frac{b z_1}{p^t z_2} & -\frac{a p^r x_2}{x_1} - \frac{b y_1}{p^s y_2} \\ & -d & c - \frac{d z_1}{p^t z_2} & -\frac{c p^r x_2}{x_1} - \frac{d y_1}{p^s y_2} \\ -1 & -\frac{p^r x_2}{x_1} & -\frac{p^r x_2 z_1}{p^t x_1 z_2} - \frac{y_1}{p^s y_2} & -\frac{p^r x_2 y_1}{p^s x_1 y_2} \end{bmatrix} \\
&\sim \begin{bmatrix} & & 1 & \\ 1 & & \frac{y_4}{y_2} & \\ & & & 1 \\ & -1 & -x_4 & -\frac{y_4}{y_2} \end{bmatrix} \\
&\quad \begin{bmatrix} & -b & a - \frac{b z_1}{p^t z_2} & -\frac{a p^r x_2}{x_1} - \frac{b y_1}{p^s y_2} - \frac{y_4}{y_2} \\ 1 & \frac{p^r x_2}{x_1} + \frac{d y_4}{y_2} & -\frac{c y_4}{y_2} + \frac{p^r x_2 z_1}{p^t x_1 z_2} + \frac{d y_4 z_1}{p^t y_2 z_2} + \frac{y_1}{p^s y_2} & \frac{c p^r x_2 y_4}{x_1 y_2} - x_4 + \frac{p^r x_2 y_1}{p^s x_1 y_2} + \frac{d y_1 y_4}{p^s y_2^2} \\ & & & 1 \\ & -d & c - \frac{d z_1}{p^t z_2} & -\frac{c p^r x_2}{x_1} - \frac{d y_1}{p^s y_2} \end{bmatrix} \\
&\sim \begin{bmatrix} & & 1 & \\ 1 & & \frac{y_4}{y_2} & \\ & & & 1 \\ & -1 & -\frac{x_4}{y_2} & -\frac{y_4}{y_2} \end{bmatrix} \begin{bmatrix} & -b & a_1 & -a_1 d_5 x_2 p^r \\ 1 & d_5 x_2 p^r & & \\ & & & 1 \\ & -d & c_1 & -c_1 d_5 x_2 p^r \end{bmatrix} \sim s_1 s_2 \begin{bmatrix} 1 & & y' & \\ & 1 & x' & y' \\ & & 1 & \\ & & & 1 \end{bmatrix},
\end{aligned}
$$

with

$$
x' = \frac{x_4}{y_2} \quad \text{and} \quad y' = \frac{y_4}{y_2}.
$$

This completes the proof in the case when both y and z are nonzero.

Case 2: $\boldsymbol{y = 0}$, $\boldsymbol{z \neq 0}$. If $y = 0$, $z \neq 0$, then we set $y_1 = 0$, $y_2 = 1$ and $s = 0$. It is easy to see that the previous proof remains valid for this case as well.

Case 3: $y \neq 0$, $z = 0$. If $z = 0$, $y \neq 0$ then we set $z_1 = 0$, $z_2 = 1$ and $t = 0$; and it is easy to see that the proof given in the first case remains valid for this case as well.

Case 4: $y = 0, z = 0$. Finally, we consider the case when both y and z are zero. In this case we make the following choices. Let

$$c = x_1, \quad d = -p^n$$

and select integers a and b such that $ad - bc = 1$. If $r \geq n$, then set

$$y_4 = \frac{x_2 p^{r-n}}{x_1}, \quad x_4 = x_2 y_4 p^r, \quad e = b x_2 p^{r-n}, \quad f = 0.$$

Otherwise if $r < n$ then set

$$y_4 = \frac{-a x_2 p^r}{x_1}, \quad x_4 = x_2 y_4 p^r, \quad e = 0, \quad f = -b x_2 p^r.$$

Now we have

$$\begin{aligned}
g &\sim s_2 s_1 s_2 s_1 s_2 \begin{bmatrix} 1 & x^{-1} & & \\ & 1 & & \\ & & 1 & -x^{-1} \\ & & & 1 \end{bmatrix} \qquad \text{(from (5-1))} \\
&\sim \begin{bmatrix} 1 & & & \\ & a & b & \\ & c & d & \\ & & & 1 \end{bmatrix} \begin{bmatrix} & & & 1 \\ & & 1 & -\frac{p^r x_2}{x_1} \\ & -1 & & \\ -1 & -\frac{p^r x_2}{x_1} & & \end{bmatrix} \\
&= \begin{bmatrix} & & 1 & \\ 1 & & y_4 & \\ & & & 1 \\ & -1 & -x_4 & -y_4 \end{bmatrix} \begin{bmatrix} & -b & a & -\frac{a p^r x_2}{x_1} - y_4 \\ 1 & d y_4 + \frac{p^r x_2}{x_1} & -c y_4 & \frac{c p^r x_2 y_4}{x_1} - x_4 \\ & & & 1 \\ & -d & c & -\frac{c p^r x_2}{x_1} \end{bmatrix} \\
&= \begin{bmatrix} & & 1 & \\ 1 & & y_4 & \\ & & & 1 \\ & -1 & -x_4 & -y_4 \end{bmatrix} \begin{bmatrix} & -b & a & e \\ 1 & f & -p^{-n+r} x_2 & \\ & & & 1 \\ & p^n & x_1 & -p^r x_2 \end{bmatrix} \\
&\sim \begin{bmatrix} & & 1 & \\ 1 & & y_4 & \\ & & & 1 \\ & -1 & -x_4 & -y_4 \end{bmatrix} \sim s_1 s_2 \begin{bmatrix} 1 & & y' & \\ & 1 & x' & y' \\ & & 1 & \\ & & & 1 \end{bmatrix}
\end{aligned}$$

with $x' = x_4$ and $y' = y_4$. This completes the proof of the lemma. □

Lemma 5.3. *Assume n and s to be positive integers. Also let p be a prime number and η be an integer such that* $\gcd(\eta, p) = 1$. *Then*

$$Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ & p^s & & 1 \end{bmatrix} \Gamma_0(p^n) = Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ \eta p^s & & 1 & \\ & \eta p^s & & 1 \end{bmatrix} \Gamma_0(p^n).$$

Proof.

If $s \geq n$, then of course both sides equal $Q(\mathbb{Q})1\Gamma_0(p^n)$. Therefore, in the following we assume that $s < n$. Next we consider the case when $n \geq 2s$ and make the following choices. Since $\gcd(\eta, p) = 1$, there exist integers α_1 and β_1 such that $\alpha_1\eta + \beta_1 p^{n-s} = 1$. Further, we also have $\gcd(\alpha_1, p^{n-s}) = 1$, so there exist β_2' and β_3' such that $\alpha_1\beta_2' + \beta_3' p^{n-s} = 1$. Let $\beta_2 = \beta_1\beta_2'$ and $\beta_3 = -\beta_1\beta_3'$. Now set

$$a = \frac{1-\beta_1 p^{n-s}}{\eta} = \alpha_1, \quad b = p^{n-2s}\beta_3, \quad c = p^n, \quad d = \eta + \beta_2 p^{n-s}.$$

We also check that

$$\begin{aligned} ad - bc &= \alpha_1(\eta + \beta_2 p^{n-s}) - \beta_3 p^{2n-2s} \\ &= 1 - \beta_1 p^{n-s} + \alpha_1\beta_2 p^{n-s} - \beta_3 p^{2n-2s} \\ &= 1 - \beta_1 p^{n-s}(1 - \alpha_1\beta_2' - \beta_3' p^{n-s}) \\ &= 1. \end{aligned}$$

On the other hand if $2s > n$, then we make the following choices. Since $\gcd(\eta, p) = 1$, there exist integers α_1 and β_1 such that $\alpha_1\eta + \beta_1 p^{n-s} = 1$. Further, we also have $\gcd(\alpha_1, p^s) = 1$, so there exist β_2' and β_3' such that $\alpha_1\beta_2' + \beta_3' p^s = 1$. Let $\beta_2 = \beta_1\beta_2'$ and $\beta_3 = -\beta_1\beta_3'$. Now set

$$a = \frac{1-\beta_1 p^{n-s}}{\eta} = \alpha_1, \quad b = \beta_3, \quad c = p^n \quad d = \eta + \beta_2 p^{n-s}.$$

Next we note

$$\begin{aligned} ad - bc = \alpha_1(\eta + \beta_2 p^{n-s}) - \beta_3 p^n &= 1 - \beta_1 p^{n-s} + \alpha_1\beta_2 p^{n-s} - \beta_3 p^n \\ &= 1 - \beta_1 p^{n-s} + \alpha_1\beta_1\beta_2' p^{n-s} + \beta_1\beta_3' p^n \\ &= 1 - \beta_1 p^{n-s}(1 - \alpha_1\beta_2' - \beta_3' p^{n-s}) \\ &= 1. \end{aligned}$$

Now the lemma follows from the following calculations:

$$
\begin{aligned}
&Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\p^s&&1&\\&p^s&&1\end{bmatrix}\Gamma_0(p^n)\\
&=Q(\mathbb{Q})\begin{bmatrix}1&&&\\&a&b&\\&c&d&\\&&&1\end{bmatrix}\begin{bmatrix}1&&&\\&1&&\\p^s&&1&\\&p^s&&1\end{bmatrix}\Gamma_0(p^n)\\
&=Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\\eta p^s&&1&\\&\eta p^s&&1\end{bmatrix}\begin{bmatrix}1&&&\\bp^s&a&b&\\dp^s-\eta p^s&c&d&\\-b\eta(p^s)^2&-a\eta p^s+p^s&-b\eta p^s&1\end{bmatrix}\Gamma_0(p^n)\\
&=Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\\eta p^s&&1&\\&\eta p^s&&1\end{bmatrix}\Gamma_0(p^n).
\end{aligned}
$$

□

Lemma 5.4. *Assume s, r and n to be positive integers with $0 < s \leq n$. Also let p be a prime number and η_1, η_2 be integers such that* $\gcd(\eta_1, p) = \gcd(\eta_2, p) = 1$. *Then*

$$
Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\p^s&&1&\\\eta_2 p^r&p^s&&1\end{bmatrix}\Gamma_0(p^n)=Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\\eta_1 p^s&&1&\\\eta_2 p^r&\eta_1 p^s&&1\end{bmatrix}\Gamma_0(p^n).
$$

Proof.
Let

$$
z_1=\begin{bmatrix}1&&&\\&1&&\\p^s&&1&\\\eta_2 p^r&p^s&&1\end{bmatrix}\quad\text{and}\quad z_2=\begin{bmatrix}1&&&\\&1&&\\\eta_1 p^s&&1&\\\eta_2 p^r&\eta_1 p^s&&1\end{bmatrix}.
$$

Then we note that

$$
z_2^{-1}\begin{bmatrix}1&&&\\&a&b&\\&c&d&\\&&&1\end{bmatrix}z_1=\begin{bmatrix}1&&&\\bp^s&a&b&\\dp^s-p^s\eta_1&c&d&\\-b(p^s)^2\eta_1&-ap^s\eta_1+p^s&-bp^s\eta_1&1\end{bmatrix}.
$$

Now, the result follows by proceeding as in the proof of Lemma 5.3. □

Lemma 5.5. *Assume n to be a positive integer and s to be a nonnegative integer. Also let p be a prime number. Let $y_1, y_2 \in \mathbb{Z}$ such that p, y_1, y_2 are pairwise coprime. Then we have the following results.*

(1) *If $s < n$ then there exists an integer b_1 with* $\gcd(b_1, p) = 1$, *such that*

$$Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ \frac{y_1}{y_2}p^s & & 1 & \\ & \frac{y_1}{y_2}p^s & & 1 \end{bmatrix}\Gamma_0(p^n) = Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ b_1 p^s & & 1 & \\ & b_1 p^s & & 1 \end{bmatrix}\Gamma_0(p^n).$$

(2) *If $s \geq n$ then*

$$Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ \frac{y_1}{y_2}p^s & & 1 & \\ & \frac{y_1}{y_2}p^s & & 1 \end{bmatrix}\Gamma_0(p^n) = Q(\mathbb{Q})1\Gamma_0(p^n).$$

Proof. Let α_1 and β_1 be integers such that

$$\alpha_1 p^s y_1 + \beta_1 y_2 = 1.$$

If $0 < s < n$ then set:

$$\alpha = \alpha_1, \quad \beta = \beta_1, \quad b_1, b_2 \in \mathbb{Z} \quad \text{such that } b_1\beta + b_2 p^{n-s} = y_1, \ b = b_1 p^s.$$

Otherwise, if $s \geq n$ then set:

$$\alpha = \alpha_1, \quad \beta = \beta_1, \quad b_2 = y_1 p^{s-n}, \quad b_1 = 0, \quad b = 0.$$

If $s = 0$ then set:

$$\alpha = \begin{cases} \alpha_1 - y_2 & \text{if } p \mid \beta_1, \\ \alpha_1 & \text{if } p \nmid \beta_1, \end{cases} \qquad \beta = \begin{cases} \beta_1 + y_1 & \text{if } p \mid \beta_1, \\ \beta_1 & \text{if } p \nmid \beta_1, \end{cases}$$

$b_1, b_2 \in \mathbb{Z}$ such that $b_1\beta + b_2 p^n = y_1$, $b = b_1$. We note that, for each of the cases considered above, i.e., whenever $s \geq 0$, the following holds:

$$-b\beta + p^s y_1 = (-b_1\beta + y_1)p^s = b_2 p^n.$$

Then,

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ \frac{y_1}{y_2}p^s & & 1 & \\ & \frac{y_1}{y_2}p^s & & 1 \end{bmatrix} \sim \begin{bmatrix} y_2^{-1} & & -\alpha & \\ & y_2^{-1} & & -\alpha \\ & & y_2 & \\ & & & y_2 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 1 & & \\ \frac{y_1}{y_2}p^s & & 1 & \\ & \frac{y_1}{y_2}p^s & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \beta & & -\alpha & \\ & \beta & & -\alpha \\ p^s y_1 & & y_2 & \\ & p^s y_1 & & y_2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & & & \\ & 1 & & \\ b & & 1 & \\ & b & & 1 \end{bmatrix} \begin{bmatrix} \beta & & -\alpha & \\ & \beta & & -\alpha \\ -b\beta + p^s y_1 & & \alpha b + y_2 & \\ & -b\beta + p^s y_1 & & \alpha b + y_2 \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & & & \\ & 1 & & \\ b_1 p^s & & 1 & \\ & b_1 p^s & & 1 \end{bmatrix}.$$

This completes the proof of the lemma. □

Lemma 5.6. *Assume n to be a positive integer, r to be a nonnegative integer and p to be a prime number. Let $x_1, x_2 \in \mathbb{Z}$ such that p, x_1, x_2 are pairwise coprime. Then we have the following results.*

(1) *If $r < n$, then there exists an integer c_1 with* $\gcd(c_1, p) = 1$ *such that*

$$Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ \frac{x_1}{x_2}p^r & & & 1 \end{bmatrix} \Gamma_0(p^n) = Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ c_1 p^r & & & 1 \end{bmatrix} \Gamma_0(p^n).$$

(2) *If $r \geq n$, then*

$$Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ \frac{x_1}{x_2}p^r & & & 1 \end{bmatrix} \Gamma_0(p^n) = Q(\mathbb{Q}) 1 \Gamma_0(p^n).$$

Proof. Let α_1 and β_1 be integers such that $\alpha_1 p^r x_1 + \beta_1 x_2 = 1$. If $0 < r < n$ then set:

$$\alpha = \alpha_1, \quad \beta = \beta_1, \quad c_1, c_2 \in \mathbb{Z} \quad \text{such that} \quad c_1\beta + c_2 p^{n-r} = x_1, \; c = c_1 p^r.$$

Otherwise, if $r \geq n$, then set:

$$\alpha = \alpha_1, \quad \beta = \beta_1, \quad c_2 = x_1 p^{r-n}, \quad c_1 = 0, \quad c = 0.$$

If $r=0$ then set:

$$\alpha=\begin{cases}\alpha_1-x_2 & \text{if } p\mid\beta_1,\\ \alpha_1 & \text{if } p\nmid\beta_1,\end{cases}\qquad \beta=\begin{cases}\beta_1+x_1 & \text{if } p\mid\beta_1,\\ \beta_1 & \text{if } p\nmid\beta_1,\end{cases}$$

$c_1,c_2\in\mathbb{Z}$ such that $c_1\beta+c_2p^n=x_1,\ c=c_1$. We note that for $r\geq 0$,

$$-c\beta+p^rx_1=(-c_1\beta+x_1)p^r=c_2p^n.$$

Then

$$\begin{aligned}\begin{bmatrix}1&&&\\&1&&\\&&1&\\\frac{x_1}{x_2}p^r&&&1\end{bmatrix}&\sim\begin{bmatrix}x_2^{-1}&&&-\alpha\\&1&&\\&&1&\\&&&x_2\end{bmatrix}\begin{bmatrix}1&&&\\&1&&\\&&1&\\\frac{x_1}{x_2}p^r&&&1\end{bmatrix}\\&=\begin{bmatrix}\beta&&&-\alpha\\&1&&\\&&1&\\p^rx_1&&&x_2\end{bmatrix}\\&=\begin{bmatrix}1&&&\\&1&&\\&&1&\\c&&&1\end{bmatrix}\begin{bmatrix}\beta&&&-\alpha\\&1&&\\&&1&\\-\beta c+p^rx_1&&&\alpha c+x_2\end{bmatrix}\sim\begin{bmatrix}1&&&\\&1&&\\&&1&\\c_1p^r&&&1\end{bmatrix}.\end{aligned}$$

This completes the proof of the lemma. □

Lemma 5.7. *Assume n to be a positive integer and p to be a prime number. Let x, y be nonzero integers coprime to p. Then we have the following results.*

(1)

$$Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\&&1&\\x&&&1\end{bmatrix}\Gamma_0(p^n)=Q(\mathbb{Q})s_1s_2\Gamma_0(p^n).$$

(2)

$$Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\y&&1&\\&y&&1\end{bmatrix}\Gamma_0(p^n)=Q(\mathbb{Q})s_1s_2\Gamma_0(p^n).$$

(3)

$$Q(\mathbb{Q})\begin{bmatrix}1&&&\\&1&&\\y&&1&\\x&y&&1\end{bmatrix}\Gamma_0(p^n)=Q(\mathbb{Q})s_1s_2\Gamma_0(p^n).$$

Proof. Let k_1 and k_2 be integers such that $k_1x + k_2p^n = 1$. Then we have

$$\begin{bmatrix}1 & & & \\ & 1 & & \\ & & 1 & \\ x & & & 1\end{bmatrix} \sim \begin{bmatrix}1 & & & \\ & 1 & & \\ & & 1 & \\ x & & & 1\end{bmatrix}\begin{bmatrix} & -k_1 & 1 & \\ 1 & & & \\ & & & 1 \\ & k_1x-1 & -x & \end{bmatrix} = \begin{bmatrix}1 & & & k_1 \\ & 1 & & \\ & & 1 & \\ & & & 1\end{bmatrix} s_1s_2 \sim s_1s_2.$$

This completes the proof of the first part of lemma.

Now, let l_1 and l_2 be integers such that $l_1y + l_2p^n = -1$. Then,

$$\begin{bmatrix}1 & & & \\ & 1 & & \\ y & & 1 & \\ & y & & 1\end{bmatrix} \sim \begin{bmatrix}1 & & & \\ & 1 & & \\ y & & 1 & \\ & y & & 1\end{bmatrix}\begin{bmatrix} l_1 & & 1 & \\ & l_1 & & 1 \\ -l_1y-1 & & -y & \\ & -l_1y-1 & & -y\end{bmatrix}$$

$$= s_2\begin{bmatrix}1 & l_1 & & \\ & 1 & & \\ & & 1 & -l_1 \\ & & & 1\end{bmatrix} s_1s_2 \sim s_1s_2.$$

This completes the proof of the second part of lemma. Finally we have

$$\begin{bmatrix}1 & & & \\ & 1 & & \\ y & & 1 & \\ x & y & & 1\end{bmatrix} \sim \begin{bmatrix}1 & & & \\ & 1 & & \\ y & & 1 & \\ x & y & & 1\end{bmatrix}\begin{bmatrix} l_1 & & 1 & \\ l_1^2x & l_1 & & 1 \\ -l_1y-1 & & -y & \\ -l_1^2xy-l_1x & -l_1y-1 & -x & -y\end{bmatrix}$$

$$= \begin{bmatrix}1 & & & \\ & l_1^2x & 1 & \\ & -1 & & \\ & & & 1\end{bmatrix}\begin{bmatrix}1 & l_1 & & \\ & 1 & & \\ & & 1 & -l_1 \\ & & & 1\end{bmatrix} s_1s_2 \sim s_1s_2.$$

This completes the proof of the last part of lemma. □

Lemma 5.8. *Assume n and r to be integers such that $0 < r < n$. Let p be a prime number and $x, y \in \mathbb{Z}$ such that $\gcd(x, p) = \gcd(y, p) = 1$. Let*

$$g_1(x, p, r) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ p^rx & & & 1\end{bmatrix} \quad \textit{and} \quad g_1(y, p, r) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ p^ry & & & 1\end{bmatrix}.$$

Then

$$Q(\mathbb{Q})g_1(x, p, r)\Gamma_0(p^n) = Q(\mathbb{Q})g_1(y, p, r)\Gamma_0(p^n),$$

if and only if

$$x \equiv y \mod p^f$$

where $f = \min(r, n-r)$.

Proof. It is clear that $g_1(x,p,r)\Gamma_0(p^n) \sim g_1(y,p,r)\Gamma_0(p^n)$ if and only if there exists an element

$$q = \begin{bmatrix} t & & & \\ & a & b & \\ & c & d & \\ & & & \frac{1}{t} \end{bmatrix} \begin{bmatrix} 1 & l & \mu & k \\ & 1 & & \mu \\ & & 1 & -l \\ & & & 1 \end{bmatrix} \in Q(\mathbb{Q}),$$

such that $g_1(y,p,r)^{-1} q g_1(x,p,r) \in \Gamma_0(p^n)$. We have

$$\begin{aligned} &g_1(y,p,r)^{-1} q g_1(x,p,r) \\ &\quad = \begin{bmatrix} kp^r tx + t & lt & \mu t & kt \\ -(bl - a\mu)p^r x & a & b & -bl + a\mu \\ -(dl - c\mu)p^r x & c & d & -dl + c\mu \\ -(kp^r ty - \frac{1}{t})p^r x - p^r ty & -lp^r ty & -\mu p^r ty & -kp^r ty + \frac{1}{t} \end{bmatrix}. \end{aligned}$$

Suppose $g_1(y,p,r)^{-1} q g_1(x,p,r) \in \Gamma_0(p^n)$. Then we must have $t = \pm 1$. We also need the condition that

$$\begin{aligned} -\Big(kp^r ty - \frac{1}{t}\Big)p^r x - p^r ty \equiv 0 \mod p^n &\Longrightarrow t^2 y - x \equiv 0 \mod p^f \\ &\Longrightarrow y - x \quad \equiv 0 \mod p^f. \end{aligned}$$

Conversely, we show that if $y - x \equiv 0 \mod p^f$, then $g_1(x,p,r)$ and $g_1(y,p,r)$ lie in the same double coset. Suppose $x - y = k_2 p^f$. As $\gcd(p^{n-r-f}, xyp^{r-f}) = 1$, there exist integers k and k_2 such that $kxyp^{r-f} + k_1 p^{n-r-f} = k_2$. So we obtain

$$-(kp^r y - 1)p^r x - p^r y = k_1 p^n.$$

Therefore

$$\begin{aligned} &g_1(y,p,r)^{-1} \begin{bmatrix} 1 & & & k \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} g_1(x,p,r) \\ &\qquad = \begin{bmatrix} kp^r x + 1 & & & k \\ & 1 & & \\ & & 1 & \\ -(kp^r y - 1)p^r x - p^r y & & & -kp^r y + 1 \end{bmatrix} \in \Gamma_0(p^n). \end{aligned}$$

This means that $g_1(x,p,r)$ and $g_1(y,p,r)$ lie in the same double coset. This completes the proof of the lemma. $\square$

Lemma 5.9. *Assume s, r and n to be integers such that $n \geq 1$, $0 < s < n$. Let p be a prime number and $x, y \in \mathbb{Z}$ such that $\gcd(x, p) = \gcd(y, p) = 1$. Let*

$$g_3(p, x, r, s) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ xp^r & p^s & & 1 \end{bmatrix}, \quad g_2(p, s) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ & p^s & & 1 \end{bmatrix}$$

$$g_3(p, y, r, s) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ yp^r & p^s & & 1 \end{bmatrix}.$$

(1) *If $r < n$ and $0 < s < r < 2s$ and $f = \min(2s - r, n - r)$, then*

$$Q(\mathbb{Q})g_3(p, x, r, s)\Gamma_0(p^n) = Q(\mathbb{Q})g_3(p, y, r, s)\Gamma_0(p^n) \iff x \equiv y \mod p^f.$$

(2) *If $2s \leq r$, then*

$$Q(\mathbb{Q})g_3(p, x, r, s)\Gamma_0(p^n) = Q(\mathbb{Q})g_2(p, s)\Gamma_0(p^n).$$

(3) *If $r \geq n$, then*

$$Q(\mathbb{Q})g_3(p, x, r, s)\Gamma_0(p^n) = Q(\mathbb{Q})g_2(p, s)\Gamma_0(p^n).$$

Proof. It is clear that $Q(\mathbb{Q})g_3(p, x, r, s)\Gamma_0(p^n) = Q(\mathbb{Q})g_3(p, y, r, s)\Gamma_0(p^n)$ if and only if there exists an element

$$q = \begin{bmatrix} t & & & \\ & a & b & \\ & c & d & \\ & & & 1/t \end{bmatrix} \begin{bmatrix} 1 & l & \mu & k \\ & 1 & & \mu \\ & & 1 & -l \\ & & & 1 \end{bmatrix} \in Q(\mathbb{Q})$$

such that $g_3(p, y, r, s)^{-1} q g_3(p, x, r, s) \in \Gamma_0(p^n)$. Suppose

$$g_3(p, y, r, s)^{-1} q g_3(p, x, r, s) \in \Gamma_0(p^n).$$

Then, comparing the multiplier of the matrices on both sides, we see that $ad - bc = 1$. Then by writing the matrix on the left explicitly it also follows that $t = \pm 1$. We can assume that $t = 1$. Now,

$$
\begin{aligned}
&g_3(p, y, r, s)^{-1} q g_3(p, x, r, s) \\
&= \begin{bmatrix}
kp^r x + \mu p^s + 1 & kp^s + l & \mu & k \\
-(bl - a\mu) p^r x + bp^s & -(bl - a\mu) p^s + a & b & -bl + a\mu \\
-(dl - c\mu + kp^s) p^r x - (\mu p^s - d) p^s - p^s & -(dl - c\mu + kp^s) p^s - lp^s + c & -\mu p^s + d & -dl + c\mu - kp^s \\
(blp^s - a\mu p^s - kp^r y + 1) p^r x - (\mu p^r y + bp^s) p^s - p^r y & -lp^r y + (blp^s - a\mu p^s - kp^r y + 1) p^s - ap^s & -\mu p^r y - bp^s & blp^s - a\mu p^s - kp^r y + 1
\end{bmatrix},
\end{aligned}
$$

and then looking at the lowest left entry we get

$$
\begin{aligned}
(blp^s - a\mu p^s - kp^r y + 1) p^r x - (\mu p^r y + bp^s) p^s - p^r y &\equiv 0 \mod p^n \\
\Longrightarrow p^r (x - y) + (bl - a\mu) x p^{r+s} - kxyp^{2r} - \mu y p^{r+s} + bp^{2s} &\equiv 0 \mod p^n \\
\Longrightarrow x - y + (bl - a\mu) x p^s - kxyp^r - \mu y p^s + bp^{2s-r} &\equiv 0 \mod p^{n-r} \\
\Longrightarrow x - y &\equiv 0 \mod p^f.
\end{aligned}
$$

Conversely, we show that if $y - x \equiv 0 \mod p^f$ then $g_3(p, x, r, s)$ and $g_3(p, y, r, s)$ lie in the same double-coset. If $f = n - r$, then let $x - y = k_1 p^{n-r}$.

$$
\begin{aligned}
Q(\mathbb{Q}) g_3(p, x, r, s) \Gamma_0(p^n) &= Q(\mathbb{Q}) g_3(p, y, r, s) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ p^r x - p^r y & & & 1 \end{bmatrix} \Gamma_0(p^n) \\
&= Q(\mathbb{Q}) g_3(p, y, r, s) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ p^n k_1 & & & 1 \end{bmatrix} \Gamma_0(p^n) \\
&= Q(\mathbb{Q}) g_3(p, y, r, s) \Gamma_0(p^n).
\end{aligned}
$$

On the other hand if $f = 2s - r$ or equivalently $2s \le n$, then let $x - y = k_2 p^{2s-r}$.

$$
\begin{aligned}
Q(\mathbb{Q}) g_3(p, x, r, s)\Gamma_0(p^n) &= Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & k_2 & \\ & & 1 & \\ & & & 1 \end{bmatrix} g_3(p, x, r, s)\Gamma_0(p^n) \\
&= Q(\mathbb{Q}) g_3(p, y, r, s) \begin{bmatrix} 1 & & & \\ k_2 p^s & 1 & k_2 & \\ & & 1 & \\ & & -k_2 p^s & 1 \end{bmatrix} \Gamma_0(p^n) \\
&= Q(\mathbb{Q}) g_3(p, y, r, s)\Gamma_0(p^n).
\end{aligned}
$$

This means that $g_3(p, x, r, s)$ and $g_3(p, y, r, s)$ lie in the same double coset and the first part of lemma follows. Next,

$$
\begin{aligned}
Q(\mathbb{Q}) g_3(p, x, r, s)\Gamma_0(p^n) &= Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & p^{r-2s}x & \\ & & 1 & \\ & & & 1 \end{bmatrix} g_3(p, x, r, s)\Gamma_0(p^n) \\
&= Q(\mathbb{Q}) g_2(p, s) \begin{bmatrix} 1 & & & \\ p^{r-2s} p^s x & 1 & p^{r-2s}x & \\ & & 1 & \\ & & -p^{r-2s} p^s x & 1 \end{bmatrix} \Gamma_0(p^n) \\
&= Q(\mathbb{Q}) g_2(p, s)\Gamma_0(p^n).
\end{aligned}
$$

This completes the proof of the second part of lemma. Finally, the last part of the lemma follows from the calculation

$$
g_2(p, s)^{-1} g_3(p, x, r, s) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ p^r x & & & 1 \end{bmatrix} \in \Gamma_0(p^n). \qquad \square
$$

Proof of Theorem 3.1.

Proof. First we prove completeness. We begin by writing

$$
\begin{aligned}
\text{(5-4)} \quad \mathrm{GSp}(4, \mathbb{Q}) = Q(\mathbb{Q}) \sqcup Q(\mathbb{Q}) s_1 &\begin{bmatrix} 1 & * & & \\ & 1 & & \\ & & 1 & * \\ & & & 1 \end{bmatrix} \\
\sqcup Q(\mathbb{Q}) s_1 s_2 &\begin{bmatrix} 1 & & * & \\ & 1 & * & * \\ & & 1 & \\ & & & 1 \end{bmatrix} \sqcup Q(\mathbb{Q}) s_1 s_2 s_1 \begin{bmatrix} 1 & * & * & * \\ & 1 & & * \\ & & 1 & * \\ & & & 1 \end{bmatrix},
\end{aligned}
$$

by using the Bruhat decomposition. We consider all the different possibilities.

First Cell: If $g \in Q(\mathbb{Q})$, then, of course, $Q(\mathbb{Q})g\Gamma_0(p^n)$ is represented by 1.

Second Cell: Assume that g is in the second cell. Then we may assume that

$$g = s_1 \begin{bmatrix} 1 & \frac{x_1}{x_2} & & \\ & 1 & & \\ & & 1 & -\frac{x_1}{x_2} \\ & & & 1 \end{bmatrix}, \quad x_1, x_2 \in \mathbb{Z} \text{ and } \gcd(x_1, x_2) = 1.$$

As $\gcd(x_1, x_2) = 1$, there exist integers l_1 and l_2 such that $-l_1x_1 + l_2x_2 = 1$.

$$\begin{aligned} Q(\mathbb{Q})g\Gamma_0(p^n) &= Q(\mathbb{Q}) \begin{bmatrix} \frac{1}{x_2} & l_1 & & \\ & x_2 & & \\ & & \frac{1}{x_2} & -l_1 \\ & & & x_2 \end{bmatrix} g\Gamma_0(p^n) \\ &= Q(\mathbb{Q}) \begin{bmatrix} l_1 & \frac{l_1x_1}{x_2} + \frac{1}{x_2} & & \\ x_2 & x_1 & & \\ & & -l_1 & \frac{l_1x_1}{x_2} + \frac{1}{x_2} \\ & & x_2 & -x_1 \end{bmatrix} \Gamma_0(p^n) = Q(\mathbb{Q})1\Gamma_0(p^n). \end{aligned}$$

Third Cell: Next let g be an element in the third cell. We may assume that

$$g = s_1s_2 \begin{bmatrix} 1 & & y & \\ & 1 & x & y \\ & & 1 & \\ & & & 1 \end{bmatrix}, \quad x, y \in Q(\mathbb{Q}).$$

The following calculation shows that we can replace x, y by $x + 1$ and $y + 1$ respectively.

$$s_1s_2 \begin{bmatrix} 1 & & y & \\ & 1 & x & y \\ & & 1 & \\ & & & 1 \end{bmatrix} \sim s_1s_2 \begin{bmatrix} 1 & & y & \\ & 1 & x & y \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} 1 & & 1 & \\ & 1 & 1 & 1 \\ & & 1 & \\ & & & 1 \end{bmatrix} \sim s_1s_2 \begin{bmatrix} 1 & & y+1 & \\ & 1 & x+1 & y+1 \\ & & 1 & \\ & & & 1 \end{bmatrix}.$$

Let x_1, x_2, x_3 and p be pairwise coprime. Also assume y_1, y_2, y_3, p to be pairwise coprime. Let $x = x_3p^{r_1}/x_2$ with $r_1 > 0$. Then the above calculation shows that we can change x to $x + 1 = (x_3p^{r_1} + x_2)/x_2$. So we can always assume x to be of the form $x_1/(x_2p^r)$ for some $r \geq 0$. Similarly, we can also assume y to be of the form $y_1/(y_2p^s)$ with $s \geq 0$. Next, suppose $\tau = \gcd(x_1, y_1) > 1$. Then replacing $x = x_1/(x_2p^r)$ by $x + \tau_1$, with τ_1 being the largest factor of y_1 that is coprime to τ, we can also assume that $\gcd(x_1, y_1) = 1$.

Now we consider all the different possibilities that may arise. First of all, it is clear that, if both x and y are in $\mathbb{Z}$, i.e., $x_2 = y_2 = 1$, $r = 0$, $s = 0$, then

$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})s_1s_2\Gamma_0(p^n)$. Next, if $x \in \mathbb{Z}$ but $y \notin \mathbb{Z}$, then

$$\begin{aligned}
Q(\mathbb{Q})g\Gamma_0(p^n) &= Q(\mathbb{Q})s_1s_2\begin{bmatrix} 1 & y & & \\ & 1 & & y \\ & & 1 & \\ & & & 1 \end{bmatrix}\Gamma_0(p^n) \\
&= Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ y & 1 & & \\ & & 1 & \\ & & -y & 1 \end{bmatrix}s_1s_2\Gamma_0(p^n) \\
&= Q(\mathbb{Q})s_1\begin{bmatrix} 1 & y^{-1} & & \\ & 1 & & \\ & & 1 & -y^{-1} \\ & & & 1 \end{bmatrix}s_1s_2\Gamma_0(p^n) \\
&= Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ y^{-1} & & 1 & \\ & y^{-1} & & 1 \end{bmatrix}\Gamma_0(p^n) = Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ \frac{y_2p^s}{y_1} & & 1 & \\ & \frac{y_2p^s}{y_1} & & 1 \end{bmatrix}\Gamma_0(p^n).
\end{aligned}$$

We note that the third equality follows from the following matrix identity:

$$\begin{bmatrix} 1 & & & \\ y & 1 & & \\ & & 1 & \\ & & -y & 1 \end{bmatrix} = \begin{bmatrix} -y^{-1} & 1 & & \\ & y & & \\ & & y^{-1} & 1 \\ & & & -y \end{bmatrix} s_1 \begin{bmatrix} 1 & y^{-1} & & \\ & 1 & & \\ & & 1 & -y^{-1} \\ & & & 1 \end{bmatrix}.$$

But, now from Lemmas 5.5 and 5.3 it follows that if $0 < s < n$, then

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ & p^s & & 1 \end{bmatrix}\Gamma_0(p^n),$$

and if $s \geq n$ then

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})1\Gamma_0(p^n).$$

Further, If $s = 0$ then from the Lemma 5.5 and 5.7 it follows that

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})s_1s_2\Gamma_0(p^n),$$

which is one of the listed representatives in the statement of the theorem. Therefore we are done in this case.

Now consider the case when $x \notin \mathbb{Z}$ and $y \in \mathbb{Z}$. Then we have

$$\begin{aligned}
Q(\mathbb{Q})g\Gamma_0(p^n) &= Q(\mathbb{Q})s_1s_2 \begin{bmatrix} 1 & & & \\ & 1 & x & \\ & & 1 & \\ & & & 1 \end{bmatrix} \Gamma_0(p^n) \\
&= Q(\mathbb{Q}) \begin{bmatrix} x & & & 1 \\ & 1 & & \\ & & 1 & \\ & & & 1/x \end{bmatrix} s_1s_2 \begin{bmatrix} 1 & & & \\ & 1 & x & \\ & & 1 & \\ & & & 1 \end{bmatrix} \Gamma_0(p^n) \\
&= Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ x^{-1} & & & 1 \end{bmatrix} \begin{bmatrix} & -1 & & \\ 1 & & & \\ & & & 1 \\ & & -1 & \end{bmatrix} \Gamma_0(p^n) \\
&= Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ x^{-1} & & & 1 \end{bmatrix} \Gamma_0(p^n).
\end{aligned}$$

Now it follows from Lemmas 5.6 and 5.7 that if $r = 0$, then

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})s_1s_2\Gamma_0(p^n),$$

and if $r \geq n$, then

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})1\Gamma_0(p^n).$$

Further, if $0 < r < n$, then Lemma 5.6 yields that

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q}) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ c_1p^r & & & 1 \end{bmatrix} \Gamma_0(p^n)$$

for some integer c_1 such that $\gcd(c_1, p) = 1$. Then it follows from Lemma 5.8 that g lies in the same double coset as one of the elements listed in the statement of the theorem.

Next, suppose $x \notin \mathbb{Z}$ and $y \notin \mathbb{Z}$. If $s = r = 0$ then from Lemmas 5.1 and 5.7 it follows that

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})s_1s_2\Gamma_0(p^n).$$

Further it follows from Lemma 5.1 that if $s \leq r$ and $r \geq n$, then

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})1\Gamma_0(p^n);$$

otherwise, if $s \le r < n$, then

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})g_1(x_3, p, r)\Gamma_0(p^n)$$

for some nonzero integer x_3 coprime to p. But then these cases have already been considered. Hence, we are left with the case when $s > r$, and then if $s \ge n$ from Lemma 5.1 we get

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})1\Gamma_0(p^n),$$

and we are done. Otherwise, still assuming $s > r$ but $s < n$, we get

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ \eta_1 p^s & & 1 & \\ \eta_2 p^{-r+2s} & \eta_1 p^s & & 1 \end{bmatrix}\Gamma_0(p^n),$$

where $\eta_1, \eta_2 \in \mathbb{Z}$ and $\gcd(\eta_i, p) = 1$ for $i = 1, 2$. In view of Lemma 5.4 it further reduces to

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})\begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ \eta_2 p^{-r+2s} & p^s & & 1 \end{bmatrix}\Gamma_0(p^n).$$

Now, the result follows from Lemma 5.9 and we are done in this case as well.

Fourth Cell: Next we consider an element g from the fourth cell and let

$$g = s_1 s_2 s_1 \begin{bmatrix} 1 & x & & \\ & 1 & & \\ & & 1 & -x \\ & & & 1 \end{bmatrix}\begin{bmatrix} 1 & & y & z \\ & 1 & & y \\ & & 1 & \\ & & & 1 \end{bmatrix}.$$

If $x \in \mathbb{Z}$ then

$$Q(\mathbb{Q})g\Gamma_0(p^n) = Q(\mathbb{Q})s_1 s_2 \begin{bmatrix} 1 & & y & \\ & 1 & z+2xy & y \\ & & 1 & \\ & & & 1 \end{bmatrix}\Gamma_0(p^n),$$

and we are reduced to the case of the third cell. Therefore let us assume that $x \notin \mathbb{Z}$. If necessary on multiplication by a suitable matrix from right, we can assume that $x = x_1/(p^r x_2)$, $y = y_1/(p^s y_2)$ and $z = z_1/(p^{r_1} z_2)$ where $x_i, y_i, z_i \in \mathbb{Z}$, for $i = 1, 2$; r, s, r_1 are nonnegative integers, x_1, x_2, p are mutually coprime integers; y_1, y_2, p are mutually coprime integers and z_1, z_2, p are also mutually coprime integers. We can further adjust x_1, y_1 and z_1 by multiplication by a proper matrix from the right, such that any two nonzero elements selected from the set $\{x_1, y_1, z_1, x_2, y_2, z_2, p\}$ are mutually coprime except, possibly, when both the chosen elements belong to

$\{x_2, y_2, z_2\}$. Then by the virtue of Lemma 5.2 once again we are reduced to the case of the third cell. This proves that the representatives listed in the theorem constitute a complete set of double coset representatives.

Disjointness. Now we prove that the double cosets represented by the representatives listed in the theorem are disjoint. It is clear that two elements w_1 and w_2 represent the same double coset if and only if there exists an element

$$q = \begin{bmatrix} t & & & \\ & a & b & \\ & c & d & \\ & & & (ad-bc)t^{-1} \end{bmatrix} \begin{bmatrix} 1 & l & \mu & k \\ & 1 & & \mu \\ & & 1 & -l \\ & & & 1 \end{bmatrix} \in Q(\mathbb{Q})$$

such that $w_2^{-1} q w_1 \in \Gamma_0(p^n)$. On comparing the multiplier on both sides we conclude that $ad - bc = 1$. Then it is also clear that we can assume $t = 1$. Also clearly q must be a matrix with integral entries. Now we consider all pairs of different representatives for checking disjointness.

$\boldsymbol{w_1 = g_3(p, \alpha, r, s),\ w_2 = g_3(p, \beta, v, w)}$: Let

$$w_1 = g_3(p,\alpha,r,s) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^s & & 1 & \\ \alpha p^r & p^s & & 1 \end{bmatrix} \quad \text{and} \quad w_2 = g_3(p,\beta,v,w) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ p^w & & 1 & \\ \beta p^v & p^w & & 1 \end{bmatrix},$$

with α, β integers coprime to p and $r, s, v, w \in \mathbb{Z}$ such that $0 < s < r < 2s$, $0 < v < w < 2v$, $0 < s,\ r < n,\ 0 < w,\ v < n$. We see that

$$w_2^{-1} q w_1 = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ -(dl - c\mu + kp^w)\alpha p^r - (\mu p^w - d)p^s - p^w & -(dl - c\mu + kp^w)p^s - lp^w + c & * & * \\ -(\beta kp^v - blp^w + a\mu p^w - 1)\alpha p^r - (\beta \mu p^v + bp^w)p^s - \beta p^v & -\beta lp^v - (\beta kp^v - blp^w + a\mu p^w - 1)p^s - ap^w & * & * \end{bmatrix}$$

Suppose $s > w$. If $w_2^{-1} q w_1 \in \Gamma_0(p^n)$, then looking at the bottom two entries of the second column we conclude that p must divide both a and c. But this contradicts that $ad - bc = 1$. Similarly, if $s < w$, by looking at first two entries of the third row we get that $p \mid d$ and $p \mid c$ contradicting $ad - bc = 1$. Therefore, we assume $s = w$. Now looking at the bottommost entry of the first column we conclude that if $r \neq v$, then the valuation of this element can not be n. Therefore, if $r \neq v$ or $s \neq w$, then $g_3(p, \alpha, r, s)$ and $g_3(p, \beta, v, w)$ lie in different double cosets. If $r = v$ and $s = w$, then Lemma 5.9 describes the condition for $g_3(p, \alpha, r, s)$ and $g_3(p, \beta, v, w)$

to lie in the same double coset. We conclude that such representatives listed in the theorem represent disjoint double cosets.

$\boldsymbol{w_1 = g_3(p, \alpha, r, s)},\ \boldsymbol{w_2 = g_2(p, w)}$: Let $w_1 = g_3(p, \alpha, r, s)$ and $w_2 = g_2(p, w)$. Assume $w_2^{-1} q w_1 \in \Gamma_0(p^n)$. Then we see that

$$w_2^{-1} q w_1 = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ -(dl - c\mu + kp^w)\alpha p^r - (\mu p^w - d)p^s - p^w & -(dl - c\mu + kp^w)p^s - lp^w + c & * & * \\ (blp^w - a\mu p^w + 1)\alpha p^r - bp^s p^w & (blp^w - a\mu p^w + 1)p^s - ap^w & * & * \end{bmatrix}$$

and it is clear that $p \mid c$ and if $s > w$ then and $p \mid a$ or else if $s < w$ then $p \mid d$. In any case $p \mid ad - bc = 1$ which is a contradiction. Hence, we further assume $s = w$. Now, as $s < r < 2s$, looking at the last entry of the first column we see that the valuation of the element $(blp^s - a\mu p^s + 1)\alpha p^r - b(p^s)^2$ is r. Since $r < n$, we conclude that $g_3(p, \alpha, r, s)$ and $g_2(p, w)$ lie in different double cosets.

$\boldsymbol{w_1 = g_3(p, \alpha, r, s), w_2 = g_1(p, \beta, v)}$: Let $w_1 = g_3(p, \alpha, r, s)$ and $w_2 = g_1(p, \beta, v)$. Assume $w_2^{-1} q w_1 \in \Gamma_0(p^n)$. Then we see that

$$w_2^{-1} q w_1 = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ -(dl - c\mu)\alpha p^r + dp^s & -(dl - c\mu)p^s + c & * & * \\ -\beta\mu p^s p^v - (\beta kp^v - 1)\alpha p^r - \beta p^v & -\beta lp^v - (\beta kp^v - 1)p^s & * & * \end{bmatrix}.$$

Clearly, $p \mid c$. Since $r > s$, p also divides d and it contradicts the condition $ad - bc = 1$. Therefore $g_3(p, \alpha, r, s)$ and $w_2 = g_1(p, \beta, v)$ lie in different double cosets.

$\boldsymbol{w_1 = g_2(p, s)},\ \boldsymbol{w_2 = g_1(p, \beta, v)}$: Let $w_1 = g_2(p, s)$ and $w_2 = g_1(p, \beta, v)$. Assume $w_2^{-1} q w_1 \in \Gamma_0(p^n)$. Then we see that

$$w_2^{-1} q w_1 = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ dp^s & -(dl - c\mu)p^s + c & * & * \\ -\beta\mu p^s p^v - \beta p^v & -\beta lp^v - (\beta kp^v - 1)p^s & * & * \end{bmatrix}.$$

Once again we see that p divides both c and d which is a contradiction to the condition $ad - bc = 1$. Therefore $g_2(p, s)$ and $w_2 = g_1(p, \beta, v)$ lie in different double cosets.

$\boldsymbol{w_1 = g_2(p, s),\ w_2 = g_2(p, w)}$: Let $w_1 = g_2(p, s)$ and $w_2 = g_1(p, w)$. Let us assume that $w_2^{-1} q w_1 \in \Gamma_0(p^n)$. Then we see that

$$w_2^{-1} q w_1 = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ dp^s & -(dl - c\mu)p^s + c & * & * \\ -\beta\mu p^s p^w - \beta p^w & -\beta l p^w - (\beta k p^w - 1)p^s & * & * \end{bmatrix}.$$

Once again we see that $p \mid c$ and if $s > w$, then $p \mid a$ or else if $s < w$, then $p \mid d$. In any case $p \mid ad - bc = 1$, which is a contradiction. Therefore $g_2(p, s)$ and $g_2(p, w)$ lie in different double cosets.

$\boldsymbol{w_1 = g_1(p, \alpha, r),\ w_2 = g_1(p, \beta, v)}$: Let $w_1 = g_1(p, \alpha, r)$ and $w_2 = g_1(p, \beta, v)$. Let us assume that $w_2^{-1} q w_1 \in \Gamma_0(p^n)$. Then we see that

$$w_2^{-1} q w_1 = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ -(dl - c\mu)\alpha p^r & c & * & * \\ -(\beta k p^v - 1)\alpha p^r - \beta p^v & -\beta l p^v & * & * \end{bmatrix}.$$

Since $r, v < n$, we see that if $r \neq v$, then valuation of $-(\beta k p^v - 1)\alpha p^r - \beta p^v$ is less than n. Therefore $g_1(p, \alpha, r)$ and $g_1(p, \beta, v)$ lie in different double cosets. This completes the proof of disjointness.

The number of representatives. Finally, we calculate the total number of inequivalent representatives. First let n be even, say $n = 2m$ for some positive integer m. Then

$$\begin{aligned} &\#(Q(\mathbb{Q}) \backslash \mathrm{GSp}(4, \mathbb{Q}) / \Gamma_0(p^{2m})) \\ &\quad = 2 + 2m - 1 + \sum_{r=1}^{2m-1} \phi(p^{\min(r, 2m-r)}) + \sum_{s=1}^{2m-1} \sum_{r=s+1}^{\min(2s-1, 2m-1)} \phi(p^{\min(2s-r, 2m-r)}) \\ &\quad = \frac{p^{m+1} + p^m - 2}{p - 1}. \end{aligned}$$

Similarly, if n is odd, say $n = 2m + 1$, then

$$\begin{aligned} &\#(Q(\mathbb{Q}) \backslash \mathrm{GSp}(4, \mathbb{Q}) / \Gamma_0(p^{2m+1})) \\ &\qquad = 2 + 2m + \sum_{r=1}^{2m} \phi(p^{\min(r, 2m+1-r)}) + \sum_{s=1}^{2m} \sum_{r=s+1}^{\min(2s-1, 2m)} \phi(p^{\min(2s-r, 2m+1-r)}) \\ &\qquad = \frac{2(p^{m+1} - 1)}{p - 1}. \end{aligned}$$

Thus on combining these we obtain the formula (3-1) for the number of one-dimensional cusps. $\square$

Proof of Lemma 3.2.

Proof. We note that the representatives for $Q(\mathbb{Q})\backslash \mathrm{GSp}(4,\mathbb{Q})/\Gamma_0(N)$ may be obtained from the representatives of $Q(\mathbb{Q})\backslash \mathrm{GSp}(4,\mathbb{Q})/\Gamma_0(p_i^{n_i})$ for $i=1$ to m. This observation is essentially based on the following two well known facts.

(1) The natural projection map from $\mathrm{Sp}(4,\mathbb{Z})$ to $\mathrm{Sp}(4,\mathbb{Z}/N\mathbb{Z})$ is surjective.

(2) $\mathrm{Sp}(4,\mathbb{Z}/\prod_p p^e\mathbb{Z}) \xrightarrow{\sim} \prod_p \mathrm{Sp}(4,\mathbb{Z}/p^e\mathbb{Z})$.

In fact, we have

$$\begin{aligned}\mathrm{Sp}(4,\mathbb{Z})/\Gamma_0(\mathrm{N}) &\xrightarrow{\sim} (\mathrm{Sp}(4,\mathbb{Z})/\Gamma(N))/(\Gamma_0(\mathrm{N})/\Gamma(N))\\ &\xrightarrow{\sim} \mathrm{Sp}(4,\mathbb{Z}/N\mathbb{Z})/\Delta(\mathbb{Z}/N\mathbb{Z}).\end{aligned}$$

Clearly, $\Delta(\mathbb{Z}/N\mathbb{Z})=\prod_{i=1}^m \Delta(\mathbb{Z}/p_i^{n_i}\mathbb{Z})$, and the following diagram is commutative.

$$\begin{array}{ccc}
\mathrm{Sp}(4,\mathbb{Z}/N\mathbb{Z}) & \xrightarrow[\psi]{\sim} & \prod_{i=1}^m \mathrm{Sp}(4,\mathbb{Z}/p_i^{n_i}\mathbb{Z})\\
\Downarrow & & \Downarrow\\
g & & (g_1,\ldots,g_m)\\
\downarrow & & \downarrow\\
g\Delta(\mathbb{Z}/N\mathbb{Z}) & & (g_1\Delta(\mathbb{Z}/p_1^{n_1}\mathbb{Z}),\ldots,g_m\Delta(\mathbb{Z}/p_m^{n_m}\mathbb{Z}))\\
\Cap & & \Cap\\
A=\mathrm{Sp}(4,\mathbb{Z}/N\mathbb{Z})/\Delta(\mathbb{Z}/N\mathbb{Z}) & \xrightarrow[\phi]{\sim} & \prod_{i=1}^m \mathrm{Sp}(4,\mathbb{Z}/p_i^{n_i}\mathbb{Z})/\Delta(\mathbb{Z}/p_i^{n_i}\mathbb{Z})=B
\end{array}$$

Next we show that the left action by $\Gamma_\infty(\mathbb{Z})$ is compatible with the isomorphisms described in the commutative diagram above. In fact, $\Gamma_\infty(\mathbb{Z})$ acts on both sides as follows:

- on A: via

$$\Gamma_\infty(\mathbb{Z})\to\Gamma_\infty(\mathbb{Z}/N\mathbb{Z}),\quad \gamma\to\bar{\gamma}$$

- on B: via

$$\begin{array}{ccccc}\Gamma_\infty(\mathbb{Z}) &\to& \Gamma_\infty(\mathbb{Z}/N\mathbb{Z}) &\xrightarrow{\sim}& \prod_{i=1}^m \Gamma_\infty(\mathbb{Z}/p_i^{n_i}\mathbb{Z})\\ \gamma &\to& \bar{\gamma} &\xrightarrow{\sim}& (\gamma_1,\gamma_2,\ldots,\gamma_{m-1},\gamma_m).\end{array}$$

Let $g\in\mathrm{Sp}(4,\mathbb{Z}/N\mathbb{Z})$, $a=g\Delta(\mathbb{Z}/N\mathbb{Z})\in A$ and $\gamma\in\Gamma_\infty(\mathbb{Z})$. Then it is easy to check that $\phi(\gamma a)=\gamma(\phi(a))$. Therefore we obtain,

$$\begin{aligned}Q(\mathbb{Q})\backslash \mathrm{GSp}(4,\mathbb{Q})/\Gamma_0(N) &\xrightarrow{\sim} (Q(\mathbb{Q})\cap\mathrm{Sp}(4,\mathbb{Z}))\backslash \mathrm{Sp}(4,\mathbb{Z})/\Gamma_0(\mathrm{N})\\ &\xrightarrow{\sim} \Gamma_\infty(\mathbb{Z}/N\mathbb{Z})\backslash \mathrm{Sp}(4,\mathbb{Z}/N\mathbb{Z})/\Delta(\mathbb{Z}/N\mathbb{Z})\\ &\xrightarrow{\sim} \prod_{i=1}^m \Gamma_\infty(\mathbb{Z}/p_i^{n_i}\mathbb{Z})\backslash \mathrm{Sp}(4,\mathbb{Z}/p_i^{n_i}\mathbb{Z})/\Delta(\mathbb{Z}/p_i^{n_i}\mathbb{Z}).\end{aligned}$$

Now the result follows from Theorem 3.1. □

Proof of Corollary 3.3.

Proof. It is easy to check that the double cosets represented by the listed representatives are disjoint. Let $\alpha(N)$ denote the total number of representatives listed in the statement of the corollary. We note that for $N = p^n$, with p a prime and $n \geq 1$, the number of listed representatives are the same as given by Theorem 3.1 (moreover, the set of representatives in this case will be seen to be equivalent to the set of representatives given by Theorem 3.1 if one applies Lemma 5.8 and Lemma 5.9 and works out the details). We will show that for any pair of coprime positive integers R and S we have $\alpha(RS) = \alpha(R)\alpha(S)$. Then it will follow that the listed representatives form a complete set because for any N their number will agree with the number given in Lemma 3.2. We have

$$\begin{aligned}
\alpha(RS) &= 1 + \sum_{\substack{\gamma \mid RS \\ 1<\gamma\leq RS}} \phi\Big(\gcd\Big(\gamma, \frac{RS}{\gamma}\Big)\Big) + \sum_{\substack{\gamma \mid RS \\ 1<\gamma\leq RS}} \sum_{\substack{\delta\mid\gamma,\, \gamma\mid\delta^2 \\ \gamma>\delta}} \phi\Big(\gcd\Big(\frac{\delta^2}{\gamma}, \frac{RS}{\gamma}\Big)\Big) \\
&= 1 + \sum_{\substack{\gamma \mid RS \\ 1<\gamma\leq RS}} \sum_{\substack{\delta\mid\gamma,\, \gamma\mid\delta^2 \\ \gamma\geq\delta}} \phi\Big(\gcd\Big(\frac{\delta^2}{\gamma}, \frac{RS}{\gamma}\Big)\Big) \\
&= 1 + \sum_{\substack{\gamma_1 \mid R,\, \gamma_2\mid S \\ 1<\gamma_1\gamma_2\leq RS}} \sum_{\substack{\delta\mid\gamma_1\gamma_2,\, \gamma_1\gamma_2\mid\delta^2 \\ \gamma_1\gamma_2\geq\delta}} \phi\Big(\gcd\Big(\frac{\delta^2}{\gamma_1\gamma_2}, \frac{RS}{\gamma_1\gamma_2}\Big)\Big) \\
&\qquad + \sum_{\substack{\gamma_1 \mid R \\ 1<\gamma_1\leq R}} \sum_{\substack{\delta\mid\gamma_1,\, \gamma_1\mid\delta^2 \\ \gamma_1\geq\delta}} \phi\Big(\gcd\Big(\frac{\delta^2}{\gamma_1}, \frac{RS}{\gamma_1}\Big)\Big) \\
&\qquad + \sum_{\substack{\gamma_2 \mid S \\ 1<\gamma_1\leq S}} \sum_{\substack{\delta\mid\gamma_2,\, \gamma_2\mid\delta^2 \\ \gamma_2\geq\delta}} \phi\Big(\gcd\Big(\frac{\delta^2}{\gamma_2}, \frac{RS}{\gamma_2}\Big)\Big) \\
&= \Bigg(1 + \sum_{\substack{\gamma_1 \mid R \\ 1<\gamma_1\leq R}} \sum_{\substack{\delta_1\mid\gamma_1,\, \gamma_1\mid\delta_1^2 \\ \gamma_1\geq\delta_1}} \phi\Big(\gcd\Big(\frac{\delta_1^2}{\gamma_1}, \frac{R}{\gamma_1}\Big)\Big)\Bigg)\Bigg(1 + \sum_{\substack{\gamma_2 \mid S \\ 1<\gamma_2\leq S}} \sum_{\substack{\delta_2\mid\gamma_2,\, \gamma_2\mid\delta_2^2 \\ \gamma_2\geq\delta_2}} \phi\Big(\gcd\Big(\frac{\delta_2^2}{\gamma_2}, \frac{S}{\gamma_2}\Big)\Big)\Bigg) \\
&= \alpha(R)\alpha(S).
\end{aligned}$$

This completes the proof.

Alternatively, instead of the above counting argument the corollary could also be proved by giving an explicit bijection between sets of representatives for $\prod_{i=1}^m Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p_i^{n_i})$ and $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(N)$. For this we recall the remark (ii) after Theorem 3.1 and note that for $m = 1$, i.e., for $N = p_1^{n_1}$, a complete set of representatives for $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(N)$ is also given by

$$g_0(\gamma, \delta, y) := \begin{bmatrix} 1 & & & \\ & 1 & & \\ \delta & & 1 & \\ y\gamma & \delta & & 1 \end{bmatrix}, \quad 1 \le \delta \le \gamma \le N,\ \gamma \mid N,\ \delta \mid N,\ \delta \mid \gamma,\ \gamma \mid \delta^2,$$

with y being the same as in the statement of the corollary.

Now suppose $N = \prod_{i=1}^{m} N_i$ with $N_i = p_i^{n_i}$ and $m > 1$. We define the map

$$\phi : \prod_{j=1}^{m} Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p_j^{n_j}) \to Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(N)$$
$$(g_0(\gamma_j, \delta_j, y_j))_{j=1,\dots,m} \to g_0(\gamma, \delta, y),$$

where γ_j and δ_j are factors of N_j such that $1 \le \delta_j \le \gamma_j \le N_j,\ \delta_j \mid \gamma_j,\ \gamma_j \mid \delta_j^2$ and $\gamma = \prod_{j=1}^{m} \gamma_j,\ \delta = \prod_{j=1}^{m} \delta_j$. Also, $y_j = L_j + \theta_j$ with $L_j = \gcd(\delta_j^2/\gamma_j, N_j/\gamma_j)$ and $\theta_j = 0$ if $L_j = 1$ otherwise $\theta_j \in (\mathbb{Z}/L_j\mathbb{Z})^\times$. Also, $y = L + \theta\beta$ with $L = \prod_{j=1}^{m} L_j$, $\beta = \prod_{p_i \nmid L,\ p_i \mid N} p_i^{n_i}$, $\theta = \sum_{j=1}^{m} \alpha_j (L/L_j)\theta_j$ and α_j is such that $\alpha_j \beta(L/L_j) \equiv 1 \mod L_j$.

It is clear that $L = \gcd(\delta^2/\gamma, N/\gamma)$. If a prime p divides L then it must divide some L_k with $1 \le k \le m$. Assume this to be the case. Then from the definition of θ it follows that $\beta\theta \equiv \theta_k \mod L_k$. As $p \mid L_k$ and $\theta_k \in (\mathbb{Z}/L_k\mathbb{Z})^\times$, it is clear that $p \nmid \theta$. Therefore $\theta \in (\mathbb{Z}/L\mathbb{Z})^\times$ as desired.

Next we show that ϕ is injective. For this let us assume that

$$\phi((g_0(\gamma_j, \delta_j, y_j))_{j=1,\dots,m}) = \phi((g_0(\gamma'_j, \delta'_j, y'_j))_{j=1,\dots,m}) = g_0(\gamma, \delta, y).$$

Then $\gamma = \prod_{j=1}^{m} \gamma_j = \prod_{j=1}^{m} \gamma'_j$ implies $\gamma_j = \gamma'_j$ for all j. Similarly $\delta_j = \delta'_j$ for all j. Hence $L_j = L'_j$ for all j. Moreover, we have $y \equiv \beta\theta \equiv \theta_j \equiv y_j \mod L_j$. Similarly $y \equiv \beta\theta \equiv \theta'_j \equiv y'_j \mod L'_j$. This gives $y_j \equiv y'_j \mod L_j$ for all j. Now Lemmas 5.8 and 5.9 imply that $g_0(\gamma_j, \delta_j, y_j)$ is equivalent to $g_0(\gamma'_j, \delta'_j, y'_j)$ in $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p_j^{n_j})$. This shows that ϕ is injective.

Finally we prove that ϕ is surjective. Assume that $g_0(\gamma, \delta, y)$ is a given representative of $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(N)$. We define γ_j and δ_j as the highest power of p_j that divides γ and δ respectively. Then we define $L_j = \gcd(\delta_j^2/\gamma_j, N_j/\gamma_j)$. Also let θ_j be defined as $\beta\theta \mod L_j$ and let $y_j = L_j + \theta_j$. It is enough to define θ_j modulo L_j because Lemmas 5.8 and 5.9 imply that if $y_j \equiv y'_j \mod L_j$ then $g_0(\gamma_j, \delta_j, y_j)$ is equivalent to $g_0(\gamma_j, \delta_j, y'_j)$ in $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p_j^{n_j})$. Hence we have uniquely defined the representative $g_0(\gamma_j, \delta_j, y_j)$ up to equivalence in $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(p_j^{n_j})$. It can be checked that $\phi((g_0(\gamma_j, \delta_j, y_j))_{j=1,\dots,m}) = g_0(\gamma, \delta, y)$ up to equivalence in $Q(\mathbb{Q})\backslash \mathrm{GSp}(4, \mathbb{Q})/\Gamma_0(N)$. Therefore ϕ is surjective and we are done. $\square$

Proof of Corollary 3.4. Since $k \geq 6$ and even the Klingen–Eisenstein series defined in the statement of the Corollary have nice convergence properties. Let $\jmath(\alpha)$ and $\jmath(\beta)$ be two one-dimensional cusps for $\Gamma_0(N)$. Let

$$E_\alpha(z) = \sum_{\jmath(\xi) \in (\jmath(\alpha) Q(\mathbb{Q}) \jmath(\alpha^{-1}) \cap \Gamma_0(N)) \backslash \Gamma_0(N)} f_\alpha(\alpha^{-1}\xi\langle Z\rangle^*) \det(j(\alpha^{-1}\xi, Z))^{-k},$$

be a Klingen–Eisenstein series associated to α. We have

$$\begin{aligned}(E_\alpha|_k\beta)(Z) &= \sum f_\alpha(\alpha^{-1}\xi\langle\beta\langle Z\rangle\rangle^*) \det(j(\alpha^{-1}\xi, \beta\langle Z\rangle))^{-k} \det(j(\beta, Z))^{-k} \\ &= \sum f_\alpha(\alpha^{-1}\xi\beta\langle Z\rangle^*) \det(j(\alpha^{-1}\xi\beta, Z))^{-k},\end{aligned}$$

where the sums are taken over $\jmath(\xi) \in (\jmath(\alpha) Q(\mathbb{Q}) \jmath(\alpha^{-1}) \cap \Gamma_0(N)) \backslash \Gamma_0(N)$. Next consider $\Phi(E_\alpha|_k\beta)(z) = \lim_{\lambda\to\infty}(E_\alpha|_k\beta)\left(\left[\begin{smallmatrix} z & \\ & i\lambda \end{smallmatrix}\right]\right)$ where Φ is the Siegel Φ operator defined earlier. The limit can be evaluated term by term because of nice convergence properties of the Eisenstein series. It follows from the proof of [Klingen 1990, Proposition 5, Chapter 5], that on taking the limit the only surviving terms are those with $\jmath(\alpha^{-1})\jmath(\xi)\jmath(\beta) \in Q(\mathbb{Q})$ with $\jmath(\xi) \in \Gamma_0(N)$. If $\jmath(\alpha)$ and $\jmath(\beta)$ are inequivalent cusps, then clearly no term survives and $\Phi(E_\alpha|_k\beta)(z) = 0$, whereas we see that $\Phi(E_\alpha|_k\alpha)(z) = f_\alpha(z)$. We have shown that each Eisenstein series is supported on a unique one-dimensional cusp. Further for a fixed one-dimensional cusp all the associated Klingen–Eisenstein series are clearly linearly independent. The corollary is now evident. □

References

[Böcherer and Ibukiyama 2012] S. Böcherer and T. Ibukiyama, "Surjectivity of Siegel Φ-operator for square free level and small weight", *Ann. Inst. Fourier (Grenoble)* **62**:1 (2012), 121–144. MR Zbl

[Diamond and Shurman 2005] F. Diamond and J. Shurman, *A first course in modular forms*, Graduate Texts in Mathematics **228**, Springer, New York, 2005. MR Zbl

[Hashimoto 1983] K.-i. Hashimoto, "The dimension of the spaces of cusp forms on Siegel upper half-plane of degree two, I", *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **30**:2 (1983), 403–488. MR Zbl

[Ibukiyama 1991] T. Ibukiyama, "On Siegel modular varieties of level 3", *Internat. J. Math.* **2**:1 (1991), 17–35. MR Zbl

[Klein 2004] M. Klein, *Verschwindungssätze für Hermitesche Modulformen sowie Siegelsche Modulformen zu den Kongruenzuntergruppen $\Gamma_0^{(n)}(N)$ und $\Gamma^{(n)}(N)$*, Ph.D. thesis, Universität des Saarlandes, 2004, Available at http://tinyurl.com/DissMKlein.

[Klingen 1990] H. Klingen, *Introductory lectures on Siegel modular forms*, Cambridge Studies in Advanced Mathematics **20**, Cambridge Univ. Press, 1990. MR Zbl

[Miyake 1989] T. Miyake, *Modular forms*, Springer, Berlin, 1989. MR Zbl

[Poor and Yuen 2007] C. Poor and D. S. Yuen, "Dimensions of cusp forms for $\Gamma_0(p)$ in degree two and small weights", *Abh. Math. Sem. Univ. Hamburg* **77** (2007), 59–80. MR Zbl

[Poor and Yuen 2013] C. Poor and D. S. Yuen, "The cusp structure of the paramodular groups for degree two", *J. Korean Math. Soc.* **50**:2 (2013), 445–464. MR Zbl

[Satake 1957/58a] I. Satake, "Compactification de espaces quotients de Siegel, II", exposé 13, 10 pp. in *Fonctions automorphes, II*, Séminaire Henri Cartan **10**, Secrétariat Mathématique, Paris, 1957/58.

[Satake 1957/58b] I. Satake, "Surjectivité globale de opérateur Φ", exposé 16, 17 pp. in *Fonctions automorphes, II*, Séminaire Henri Cartan **10**, Secrétariat Mathématique, Paris, 1957/58.

[Tsushima 2003] R. Tsushima, "Dimension formula for the spaces of Siegel cusp forms of half integral weight and degree two", *Comment. Math. Univ. St. Pauli* **52**:1 (2003), 69–115. MR Zbl

Received February 4, 2017. Revised August 5, 2017.

ALOK SHUKLA
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OKLAHOMA
NORMAN, OK
UNITED STATES
alok.shukla@ou.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 293, No. 1, 2018
dx.doi.org/10.2140/pjm.2018.293.245

NONEXISTENCE RESULTS FOR SYSTEMS OF ELLIPTIC AND PARABOLIC DIFFERENTIAL INEQUALITIES IN EXTERIOR DOMAINS OF $\mathbb{R}^n$

YUHUA SUN

We present a unified approach for the investigation of nonexistence results of systems of elliptic and parabolic differential inequalities. Our results accord with those on elliptic differential inequalities given by Bidaut-Véron and Pohozaev. The results on systems of parabolic differential inequalities are new.

1. Introduction

In this paper, we study the nonexistence of nonnegative solutions to systems of the following elliptic and parabolic differential inequalities

$$\begin{cases} \Delta u + |x|^a v^p \le 0 & \text{in } \bar{D}^c, \\ \Delta v + |x|^b u^q \le 0 & \text{in } \bar{D}^c, \\ u(x) \ge f(x), \quad v(x) \ge g(x) & \text{on } \partial D, \end{cases} \tag{1-1}$$

and

$$\begin{cases} \Delta u - \partial_t u + |x|^a v^p \le 0 & \text{in } \bar{D}^c \times (0,\infty), \\ \Delta v - \partial_t v + |y|^b u^q \le 0 & \text{in } \bar{D}^c \times (0,\infty), \\ u(x,t) \ge f(x), \quad v(x,t) \ge g(x) & \text{on } \partial D \times (0,\infty), \\ u(x,0) = u_0(x), \quad v(x,0) = v_0(x) & \text{in } \bar{D}^c, \end{cases} \tag{1-2}$$

where D is a bounded Lipschitz domain in $\mathbb{R}^n$ with $n \ge 3$ containing the origin, and $\bar{D}^c = \mathbb{R}^n \setminus \bar{D}$. The exponents satisfy $a, b > -2$ and $p, q > 1$, and $f(x)$, $g(x)$ are $L^1(\partial D)$ nonnegative and positive somewhere functions, and $u_0(x)$, $v_0(x)$ are nonnegative functions.

It is well known that the nonexistence theorems for elliptic equations started from the seminal work by Gidas and Spruck [1981], where they proved the following results for the semilinear problem

$$\Delta u + u^p = 0, \quad \text{in } \mathbb{R}^n. \tag{1-3}$$

MSC2010: primary 35K51; secondary 35M87.
Keywords: system of differential inequalities, critical exponents, exterior domain.

If

$$1 < p < \frac{n+2}{n-2}, \tag{1-4}$$

then the only nonnegative solution of (1-3) is identically zero.

In 1986, Ni and Serrin showed that the exponent $(n+2)/(n-2)$ in (1-4) is critical; namely, if $p \geq (n+2)/(n-2)$, then there exist nontrivial positive solutions to (1-3). We refer to the papers [Ni and Serrin 1986a; 1986b] for more information.

In the study of equation (1-3) in the exterior domain $\mathbb{R}^n \setminus \overline{B_1(0)}$ instead of the entire Euclidean space $\mathbb{R}^n$, some incredible phenomena arise. Here $B_r(0)$ is the ball of radius r centered at the origin. This marvelous result is due to Bidaut-Véron [1989]: if

$$1 < p \leq \frac{n}{n-2}, \tag{1-5}$$

then the only nonnegative solution of (1-3) in the exterior domain is identically zero. Actually Bidaut-Véron [1989] obtained more generalized results on the problem $\Delta_m u + u^p = 0$ in the exterior domain under additional restrictions on m and p. Here to compare with Gidas and Spruck's result profitably, we only list the nonexistence result for $m = 2$. However, if $p > n/(n-2)$, the nonexistence result does not hold any more. A simple counterexample is given by the function

$$u(x) = \lambda |x|^{-2/(p-1)},$$

which is a well-defined solution to (1-3) in $\mathbb{R}^n \setminus \overline{B_1(0)}$, where

$$\lambda = (p-1)^{-2/(p-1)} \Big[2(n-2) \Big(p - \frac{n}{n-2} \Big) \Big]^{1/(p-1)}.$$

Let us turn our attention to the elliptic differential inequality case; namely, consider the problem

$$\Delta u + u^p \leq 0, \quad \text{in } \mathbb{R}^n, \tag{1-6}$$

with $n > 2$. Ni and Serrin [1986a] proved that if

$$1 < p \leq \frac{n}{n-2}, \tag{1-7}$$

then the only nonnegative solution of (1-6) is identically zero. For more elliptic differential inequality cases, we refer to the papers [Caristi et al. 2008; 2009; Mitidieri and Pohozaev 1998; 2001].

Bidaut-Véron and Pohozaev [2001] showed that in the exterior domain $\mathbb{R}^n \setminus \overline{B_1(0)}$, if under the same condition (1-7) as in the entire Euclidean space, then the only nonnegative solution of (1-6) in the exterior domain $\mathbb{R}^n \setminus \overline{B_1(0)}$ is identically zero. It is easy to see that in the inequality case, the critical exponents arising from the entire Euclidean space and exterior domain settings are the same. The difference between the entire Euclidean space and exterior domain vanishes when we move our focus from equation to differential inequality problems.

Now, let us provide some motivations from the point of view of parabolic equations. The study of critical exponents of the parabolic equation also has a long story. When D is empty, in a celebrated paper, Fujita [1966] proved that for the problem

$$\begin{cases} \partial_t u - \Delta u = u^p & \text{in } \mathbb{R}^n \times (0,\infty), \\ u(x,0) = u_0(x) & \text{in } \mathbb{R}^n. \end{cases} \tag{1-8}$$

(1) If $1 < p < 1+2/n$ and $u_0 > 0$, then (1-8) possesses no global positive solution.

(2) If $p > 1+2/n$ and u_0 is smaller than a small Gaussian, then (1-8) has global solutions.

Usually, we call $1+2/n$ the Fujita exponent. The sharpness of $p = 1+2/n$ is more difficult. Several authors independently showed that $p = 1+2/n$ belongs to the blowup case; we refer to the papers [Aronson and Weinberger 1978; Hayakawa 1973; Kobayashi et al. 1977]. Let us replace $\mathbb{R}^n$ by $\bar{D}^c$ in (1-8) (here D is a bounded nonempty domain), and we have an additional boundary condition $u|_{\partial D} \equiv f(x) \geq 0$. If the boundary condition $f(x) \equiv 0$, Bandle and Levine [1989] proved the Fujita exponent is still $p = 1+2/n$ for (1-8). But, if the boundary condition $f(x)$ is not identically zero, Zhang found that the Fujita exponent for (1-8) will jump from $1+2/n$ to a much bigger value $1+2/(n-2)$; see [Zhang 2001].

Laptev [2003] considered the scalar case of (1-2) with the nonzero boundary condition

$$\partial_t u - \Delta u \geq |x|^a u^p, \qquad \bar{D}^c \times (0,\infty), \tag{1-9}$$

and obtained that if $1 < p < (n+1+a)/(n-1)$, then (1-9) has no nontrivial nonnegative global solutions.

Motivated by the above literature, we investigate systems of elliptic and parabolic differential inequalities. First, let us explain in which sense solutions of (1-2) are defined.

Definition 1.1. A nonnegative pair (u, v) is called a weak nonnegative global solution of the inequality system (1-2), if

(i) $\nabla_x u, \nabla_x v \in L^2_{loc}(\bar{D}^c)$;

(ii) For all compactly supported $\psi \in C^2(\bar{D}^c \times [0,\infty)) \cap C^1(D^c \times [0,\infty))$ vanishing on $\partial D \times [0,\infty)$, and for all $\tau \in [0,\infty)$, we have

$$\begin{cases} \int_0^\tau \int_{\bar{D}^c} [u\Delta\psi + u\partial_t\psi + |y|^a v^p(y,s)\psi]\,dy\,ds \\ \qquad - \int_0^\tau \int_{\partial D} f(\partial\psi/\partial n^+)\,dS_y\,ds - \int_{\bar{D}^c} u(x,\cdot)\psi(x,\cdot)|_0^\tau\,dx \leq 0, \\ \int_0^\tau \int_{\bar{D}^c} [v\Delta\psi + v\partial_t\psi + |y|^b u^q(y,s)\psi]\,dy\,ds \\ \qquad - \int_0^\tau \int_{\partial D} g(\partial\psi/\partial n^+)\,dS_y\,ds - \int_{\bar{D}^c} v(x,\cdot)\psi(x,\cdot)|_0^\tau\,dx \leq 0. \end{cases} \tag{1-10}$$

Here, n^+ means the outward unit normal of ∂D, relative to D^c, which is defined almost everywhere.

Throughout, when we say that (u, v) is a global positive solution of (1-2), we mean that $u, v \geq 0$ and $u(x,t)$, $v(x,t)$ are not identically zero for each $t > 0$.

Here are our main results:

Theorem 1.2. *Assume $p \geq q > 1$. If*

$$\max\left\{\frac{2p(1+q)+bp+a}{pq-1}, \frac{2q(1+p)+aq+b}{pq-1}\right\} > n, \tag{1-11}$$

then there exist no global positive solutions to (1-2).

Corollary 1.3. *Assume $p \geq q > 1$. If*

$$\max\left\{\frac{2p(1+q)+bp+a}{pq-1}, \frac{2q(1+p)+aq+b}{pq-1}\right\} > n, \tag{1-12}$$

then there exist no positive solutions to (1-1).

Theorem 1.2 and Corollary 1.3 require that D is not empty, since we technically depend on Proposition 2.1. Corollary 1.3 was also obtained by Bidaut-Véron and Pohozaev [2001]. We claim that our technique is quite different from the one in that work, where they investigated various elliptic inequalities, and their technique is to multiply the elliptic inequalities (1-1) by functions $u^\alpha\varphi$, $v^\beta\varphi$ and to obtain the integral estimates with respect to the polynomials of u, v near infinity, where φ has compact support in $\bar{D}^c$, and $\alpha, \beta < 0$. However, we mainly investigate the parabolic differential inequalities. As a byproduct, we obtain the same result for the elliptic problem. Our method, motivated by [Zhang 1998; 1999; 2001], is to show that the integrals of I_R, J_R in (2-9) and (2-10) will blow up in some selected fixed domain.

We also improve the result obtained by Laptev [2003]. When $u = v$, $p = q$, $a = b$, the system (1-2) is reduced to the scalar case (1-9). From Theorem 1.2, it is easy to obtain that if $1 < p < (n+a)/(n-2)$, then (1-9) admits no nontrivial nonnegative global solutions. Our exponent $(n+a)/(n-2)$ here is strictly bigger than $(n+1+a)/(n-1)$ which is obtained by Laptev. We claim that our method is also different from Laptev's, since his method is based on the test function approach, which was developed by Mitidieri and Pohozaev [1998; 2001].

Notation. The letters $C, C', C_0, C_1, \ldots$ denote positive constants whose values are unimportant and may vary at different occurrences.

2. Proof of Theorem 1.2

In this section, we show the proof of Theorem 1.2. Since every positive solution (u, v) of the elliptic problem (1-1) can also be considered as a global nontrivial positive solution of the parabolic inequality system (1-2), it suffices to show that the parabolic system (1-2) has no global positive solution, provided that

$$\max\left\{\frac{2p(1+q)+bp+a}{pq-1}, \frac{2q(1+p)+aq+b}{pq-1}\right\} > n.$$

Before presenting the proof, let us cite a result which is proved in [Zhang 2001].

Proposition 2.1. *Let $\zeta_i = \zeta_i(x,t)$, $i = 1, 2$ be the solution of the linear problem*

$$\begin{cases} \Delta\zeta - \partial_t\zeta = 0 & \text{in } \bar{D}^c \times (0,\infty), \\ \zeta(x,t) = f(x) \quad (\text{respectively } g(x)) & \text{on } \partial D \times (0,\infty), \\ \zeta(x,0) = 0 & \text{in } \bar{D}^c. \end{cases} \tag{2-1}$$

If $f(x)$, $g(x)$ are nonnegative and positive somewhere, then there exist positive constants C and R_0 such that

$$\zeta_1(x,t), \zeta_2(x,t) \geq \frac{C}{R^{n-2}}, \quad \text{if } R_0 \leq R \leq |x| \leq 2R,\ R^{4n} \leq t. \tag{2-2}$$

Now we step into the proof of Theorem 1.2.

Proof of Theorem 1.2. Let

$$\omega_1(x,t) := u(x,t) - \zeta_1(x,t), \quad \omega_2(x,t) := v(x,t) - \zeta_2(x,t), \tag{2-3}$$

From (1-2) and (2-1), we derive that ω_1, ω_2 satisfy the following problems:

$$\begin{cases} \Delta\omega_1 - \partial_t\omega_1 + |x|^a(\omega_2+\zeta_2)^p \leq 0, & \text{in } \bar{D}^c \times (0,\infty) \\ \omega_1(x,t) \geq 0, & \text{on } \partial D \times (0,\infty), \\ \omega_1(x,0) = u_0(x), & \text{in } \bar{D}^c, \end{cases} \tag{2-4}$$

and

$$\begin{cases} \Delta\omega_2 - \partial_t\omega_2 + |x|^b(\omega_1+\zeta_1)^q \leq 0, & \text{in } \bar{D}^c \times (0,\infty) \\ \omega_2(x,t) \geq 0, & \text{on } \partial D \times (0,\infty), \\ \omega_2(x,0) = v_0(x), & \text{in } \bar{D}^c. \end{cases} \tag{2-5}$$

Moreover, applying the maximum principle, we know that ω_1, ω_2 are nonnegative functions.

Since

$$|x|^a(\omega_1+\zeta_1)^q \geq |x|^a\omega_1^q + |x|^a\zeta_1^q, \quad |x|^b(\omega_2+\zeta_2)^p \geq |x|^b\omega_2^p + |x|^b\zeta_2^p,$$

we obtain that

$$\Delta\omega_1 - \partial_t\omega_1 + |x|^a\omega_2^p + |x|^a\zeta_2^p \leq 0, \tag{2-6}$$

and

$$\Delta\omega_2 - \partial_t\omega_2 + |x|^b\omega_1^q + |x|^b\zeta_1^q \leq 0. \tag{2-7}$$

Introduce two functions $\varphi, \eta \in C^\infty[0, \infty)$ which satisfy the following conditions:

(i) $0 \le \varphi \le 1$; $\quad \varphi(r) = 1,\ r \in [2, 3]$; $\quad \varphi(r) = 0,\ r \in [0, 1) \cup (4, \infty)$;

(ii) $|\varphi'(r)| \le C$; $\quad \varphi'(1) = \varphi'(4) = 0$; $\quad |\varphi''(r)| \le C$;

(iii) $0 \le \eta \le 1$; $\quad \eta(t) = 1,\ t \in \left[0, \frac{1}{4}\right]$; $\quad \eta(t) = 0,\ t \in [1, \infty)$; $\quad -C \le \eta'(t) \le 0$.

Since D is bounded, we can choose $R > 0$ large enough so that $D \subset B_R(0)$. Denote

$$\varphi_R(x) := \varphi\Big(\frac{|x|}{R}\Big), \quad \eta_R(t) := \eta\Big(\frac{t - R^{4n}}{R^2}\Big).$$

It is obvious that

$$\left|\frac{\partial \varphi_R}{\partial r}\right| \le \frac{C}{R}, \quad \left|\frac{\partial^2 \varphi_R}{\partial r^2}\right| \le \frac{C}{R^2}, \quad -\frac{C}{R^2} \le \eta'_R(t) \le 0, \tag{2-8}$$

and also

$$\frac{\partial \varphi_R(x)}{\partial r} = 0 \quad \text{for } |x| = R \text{ or } |x| = 4R.$$

Denote

$$Q_R := [B_{4R}(0) \setminus B_R(0)] \times [R^{4n}, R^{4n} + R^2],$$

and also

$$\psi_R(x, t) := \varphi_R(x)\eta_R(t).$$

Let us estimate the following two integrals:

$$I_R := \int_{Q_R} |x|^a \omega_2^p(x, t)\psi_R^{q'}(x, t)\, dx\, dt, \tag{2-9}$$

and

$$J_R := \int_{Q_R} |x|^b \omega_1^q(x, t)\psi_R^{q'}(x, t)\, dx\, dt, \tag{2-10}$$

where q' is Hölder conjugate to q, satisfying $1/q + 1/q' = 1$.

Since $\omega_1(x, t)$ is a nonnegative solution of (2-6), we obtain

$$I_R + \int_{Q_R} |x|^a \zeta_2^p \psi_R^{q'}(x, t)\, dx\, dt \le \int_{Q_R} [\partial_t \omega_1 - \Delta \omega_1]\psi_R^{q'}(x, t)\, dx\, dt,$$

Integration by parts yields

$$\begin{aligned}
I_R+\int_{Q_R}|x|^a\zeta_2^p\psi_R^{q'}(x,t)\,dx\,dt \le \int_{B_{4R(0)}\setminus B_R(0)} &\omega_1(x,\cdot)\psi_R^{q'}(x,\cdot)\Big|_{R^{4n}}^{R^{4n}+R^2}dx\\
&-q'\int_{Q_R}\omega_1(x,t)\varphi_R^{q'}(x)\eta_R^{q'-1}(t)\eta_R'(t)\,dx\,dt\\
&+\int_{R^{4n}}^{R^{4n}+R^2}\int_{\partial B_{4R}(0)}\omega_1(x,t)\frac{\partial\varphi_R^{q'}(x)}{\partial n}\eta_R^{q'}(t)\,dS_x\,dt\\
&-\int_{R^{4n}}^{R^{4n}+R^2}\int_{\partial B_{4R}(0)}\psi_R^{q'}\frac{\partial\omega_1}{\partial n}(x,t)\,dS_x\,dt\\
&-\int_{R^{4n}}^{R^{4n}+R^2}\int_{\partial B_{R}(0)}\omega_1(x,t)\frac{\partial\varphi_R^{q'}(x)}{\partial n}\eta_R^{q'}(t)\,dS_x\,dt\\
&+\int_{R^{4n}}^{R^{4n}+R^2}\int_{\partial B_{R}(0)}\psi_R^{q'}\frac{\partial\omega_1}{\partial n}(x,t)\,dS_x\,dt\\
&-\int_{Q_R}\omega_1(x,t)\Delta\varphi_R^{q'}(x)\eta_R^{q'}(t)\,dx\,dt.
\end{aligned}$$

Noting here that $-\partial/\partial n=\partial/\partial n^+,\ \omega_1(x,R^{4n})\ge 0$,

$$\frac{\partial\varphi_R^{q'}}{\partial n}=q'\varphi_R^{q'-1}\frac{\partial\varphi_R}{\partial n}=0\quad\text{on }\partial B_R(0)\cup\partial B_{4R}(0)$$

and $\psi_R(x,t)=0$ on the lateral boundary of Q_R, we obtain

$$\begin{aligned}
\text{(2-11)}\quad I_R+\int_{Q_R}|x|^a\zeta_2^p\psi_R^{q'}(x,t)dxdt \le -q'&\int_{Q_R}\omega_1(x,t)\varphi_R^{q'}(x)\eta_R^{q'-1}(t)\eta_R'(t)dxdt\\
&-\int_{Q_R}\omega_1(x,t)\Delta\varphi_R^{q'}(x)\eta_R^{q'}(t)dxdt.
\end{aligned}$$

Since $\Delta\varphi_R^{q'}=q'\varphi_R^{q'-1}\Delta\varphi_R+q'(q'-1)\varphi_R^{q'-2}|\nabla\varphi_R|^2$, combining with (2-11), we get

$$\begin{aligned}
\text{(2-12)}\quad I_R+\int_{Q_R}|x|^a\zeta_2^p\psi_R^{q'}\,dx\,dt \le -q'&\int_{Q_R}\omega_1(x,t)\varphi_R^{q'}(x)\eta_R^{q'-1}(t)\eta_R'(t)\,dx\,dt\\
-q'&\int_{Q_R}\omega_1(x,t)\varphi_R^{q'-1}(x)\Delta\varphi_R(x)\eta_R^{q'}(t)\,dx\,dt.
\end{aligned}$$

By the definition of φ_R and η_R, and applying Proposition 2.1, for large R, we obtain

$$\begin{aligned}
\int_{Q_R}|x|^a\zeta_2^p\psi_R^{q'}(x,t)\,dx\,dt &\ge \int_{R^{4n}}^{R^{4n}+R^2/4}\int_{B_{3R}(0)\setminus B_{2R}(0)}|x|^a\zeta_2^p\,dx\,dt\\
&\ge CR^{n+2+a-p(n-2)}.
\end{aligned}$$

It follows from (2-12) that

$$I_R + CR^{n+2+a-p(n-2)} \le -q' \int_{Q_R} \omega_1(x,t)\varphi_R^{q'}(x)\eta_R^{q'-1}(t)\eta_R'(t)\,dx\,dt - q' \int_{Q_R} \omega_1(x,t)\varphi_R^{q'-1}(x)\Delta\varphi_R(x)\eta_R^{q'}(t)\,dx\,dt. \tag{2-13}$$

Noting that φ_R is radial, we obtain $\Delta\varphi_R = \varphi_R'' + (n-1)/r\varphi_R'$. For large enough R,

$$|\Delta\varphi_R| \le \frac{C}{R^2}, \quad \text{for } x \in B_{4R}(0)\setminus B_R(0). \tag{2-14}$$

Combining (2-13) and (2-14), we obtain

$$\begin{aligned} I_R + CR^{n+2+a-p(n-2)} &\le \frac{C}{R^2}\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} \omega_1(x,t)\varphi_R^{q'}(x)\eta_R^{q'-1}(t)\,dx\,dt \\ &\quad + \frac{C}{R^2}\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} \omega_1(x,t)\varphi_R^{q'-1}(x)\eta_R^{q'}(t)\,dx\,dt. \end{aligned}$$

According to the assumptions that $\varphi_R(x)$, $\eta_R(t) \le 1$ and $\psi_R(x,t) = \varphi_R(x)\eta_R(t)$, we have $\varphi_R^{q'}(x)\eta_R^{q'-1}(t) \le \psi_R^{q'-1}(x,t)$, and $\varphi_R^{q'-1}(x)\eta_R^{q'}(t) \le \psi_R^{q'-1}(x,t)$. Applying the Hölder inequality to the above, we obtain

$$\begin{aligned} I_R + CR^{n+2+a-p(n-2)} &\le \frac{C}{R^2}\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^b\omega_1^q\psi_R^{q(q'-1)}(x,t)\,dx\,dt\right]^{1/q} \\ &\qquad \times \left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^{-bq'/q}\,dx\,dt\right]^{1/q'} \\ &\quad + \frac{C}{R^2}\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^b\omega_1^q\psi_R^{q(q'-1)}(x,t)\,dx\,dt\right]^{1/q} \\ &\qquad \times \left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^{-bq'/q}\,dx\,dt\right]^{1/q'}. \end{aligned}$$

Hence

$$\begin{aligned} I_R + CR^{n+2+a-p(n-2)} &\le C\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^b\omega_1^q\psi_R^{q'}(x,t)\,dx\,dt\right]^{\frac{1}{q}} R^{\frac{(n+2)(q-1)-b}{q}-2} \\ &\quad + C\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^b\omega_1^q\psi_R^{q'}(x,t)\,dx\,dt\right]^{\frac{1}{q}} R^{\frac{(n+2)(q-1)-b}{q}-2}, \end{aligned}$$

which yields

$$I_R + CR^{n+2+a-p(n-2)} \le C J_R^{\frac{1}{q}} R^{\frac{(n+2)(q-1)-b}{q}-2}, \tag{2-15}$$

where we have used the definition of J_R in (2-10).

Using the same arguments with J_R, we obtain an analogous inequality for J_R. Since $\omega_2(x,t)$ is a solution of (2-7), we have

$$J_R + \int_{Q_R} |x|^b \zeta_1^q \psi_R^{q'}(x,t)\,dx\,dt \le \int_{Q_R} [\partial_t \omega_2 - \Delta\omega_2]\psi_R^{q'}(x,t)\,dx\,dt.$$

It follows that

$$\begin{aligned} J_R + CR^{n+2+b-q(n-2)} \le \frac{C}{R^2}\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} \omega_2(x,t)\varphi_R^{q'}\eta_R^{q'-1}(t)\,dx\,dt \\ + \frac{C}{R^2}\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} \omega_2(x,t)\varphi_R^{q'-1}(x)\eta_R^{q'}(t)\,dx\,dt. \end{aligned}$$

Applying the Hölder inequality, we obtain

$$\begin{aligned} &J_R + CR^{n+2+b-q(n-2)} \\ &\le \frac{C}{R^2}\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^a \omega_2^p(x,t)\psi_R^{p(q'-1)}(x,t)\,dx\,dt\right]^{1/p} \\ &\quad\times\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^{-ap'/p}\,dx\,dt\right]^{1/p'} \\ &\quad+\frac{C}{R^2}\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^a \omega_2^p(x,t)\psi_R^{p(q'-1)}(x,t)\,dx\,dt\right]^{1/p} \\ &\quad\times\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^{-ap'/p}\,dx\,dt\right]^{1/p'}, \end{aligned}$$

where $1/p+1/p'=1$. Using that $\psi_R(x,t)\le 1$, $p\ge q$ and $\psi_R^{p(q'-1)} \le \psi_R^{q(q'-1)} = \psi_R^{q'}$, we obtain

$$\begin{aligned} J_R + CR^{n+2+b-q(n-2)} &\le \frac{C}{R^2}\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^a \omega_2^p(x,t)\psi_R^{q'}(x,t)\,dx\,dt\right]^{1/p} \\ &\quad\times\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^{-ap'/p}\,dx\,dt\right]^{1/p'} \\ &\quad+\frac{C}{R^2}\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^a \omega_2^p(x,t)\psi_R^{q'}(x,t)\,dx\,dt\right]^{1/p} \\ &\quad\times\left[\int_{R^{4n}}^{R^{4n}+R^2}\int_{B_{4R}(0)\setminus B_R(0)} |x|^{-ap'/p}\,dx\,dt\right]^{1/p'}. \end{aligned}$$

which is

$$J_R + CR^{n+2+b-q(n-2)} \le C I_R^{\frac{1}{p}} R^{\frac{(n+2)(p-1)-a}{p}-2}. \tag{2-16}$$

Case 1: If

$$\frac{2p(q+1)+bp+a}{pq-1} = \max\left\{\frac{2p(q+1)+bp+a}{pq-1}, \frac{2q(1+p)+aq+b}{pq-1}\right\} > n.$$

Combining (2-15) and (2-16), we obtain

$$J_R + C_0 R^{n+2+b-q(n-2)} \le C_1 J_R^{\frac{1}{pq}} R^{\frac{(n+2)(pq-1)-b-aq}{pq}-\frac{2(1+p)}{p}}. \tag{2-17}$$

Denote

$$\begin{aligned} k_0 &:= n+2+b-q(n-2), \\ k_1 &:= \frac{(n+2)(pq-1)-b-aq}{pq} - \frac{2(1+p)}{p}. \end{aligned} \tag{2-18}$$

From (2-17), we obtain

$$J_R \ge \left(\frac{C_0}{C_1}\right)^{pq} R^{k_0(pq)-k_1(pq)}. \tag{2-19}$$

Substituting (2-19) into the left-hand side of (2-17), we obtain

$$J_R \ge \frac{C_0^{(pq)^q}}{C_1^{pq+(pq)^2}} R^{k_0(pq)^2-k_1(pq)^2-k_1(pq)}.$$

Repeating the above procedure, we obtain for any integer $j > 1$

$$J_R \ge \frac{C_0^{(pq)^j}}{C_1^{pq+\cdots+(pq)^j}} R^{k_0(pq)^j-k_1[pq+\cdots+(pq)^j]}. \tag{2-20}$$

The exponent of R in the right-hand side of (2-20) gives

$$\begin{aligned} k_0(pq)^j - k_1[pq+\cdots+(pq)^j] &= k_0(pq)^j - k_1 pq\frac{(pq)^j-1}{pq-1} \\ &= (pq)^j\left[k_0 - \frac{k_1 pq}{pq-1}\right] + \frac{k_1 pq}{pq-1}. \end{aligned}$$

From (2-20), we obtain

$$\begin{aligned} J_R &\ge C_2^{(pq)^j} R^{(pq)^j[k_0-k_1pq/(pq-1)]} R^{k_1pq/(pq-1)} \\ &= (C_2 R^{k_0-k_1pq/(pq-1)})^{(pq)^j} R^{k_1pq/(pq-1)}. \end{aligned} \tag{2-21}$$

Combining with (2-18), we obtain

$$k_0-\frac{k_1pq}{pq-1}=n+2+b-q(n-2)-\left[\frac{(n+2)(pq-1)-b-aq}{pq}-\frac{2(1+p)}{p}\right]\frac{pq}{pq-1}$$
$$=\frac{2pq(1+q)+bpq+aq-nq(pq-1)}{pq-1}.$$

Obviously, if

$$\frac{2p(1+q)+bp+a}{pq-1}>n,$$

then $k_0-k_1pq/(pq-1)>0$. Whence, if R is chosen large enough, we have $C_2R^{k_0-k_1pq/(pq-1)}>1$.

From (2-21), for fixed R, letting $j\to\infty$, we obtain

$$J_R=\int_{Q_R}|x|^b\omega_1^q(x,t)\psi_R^{q'}(x,t)\,dx\,dt=\infty. \tag{2-22}$$

However, the above contradicts (2-17), since (2-17) implies that $J_R\le CR^{k_1pq/(pq-1)}$. Moreover, by the definition of J_R, (2-22) means that $u(x,t)$ has to blow up when $t\le R^{4n}+R^2$.

Case 2: If

$$\frac{2q(1+p)+aq+b}{pq-1}=\max\left\{\frac{2p(q+1)+bp+a}{pq-1},\frac{2q(1+p)+aq+b}{pq-1}\right\}>n,$$

one can argue in the same way as with I_R and obtain the same contradiction. Hence, we finish the proof. □

Acknowledgments

We would like to thank the referees very much for their valuable comments and suggestions. This research was in part supported by the National Natural Science Foundation of China (No. 11501303), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[Aronson and Weinberger 1978] D. G. Aronson and H. F. Weinberger, "Multidimensional nonlinear diffusion arising in population genetics", *Adv. in Math.* **30**:1 (1978), 33–76. MR Zbl

[Bandle and Levine 1989] C. Bandle and H. A. Levine, "On the existence and nonexistence of global solutions of reaction-diffusion equations in sectorial domains", *Trans. Amer. Math. Soc.* **316**:2 (1989), 595–622. MR Zbl

[Bidaut-Véron 1989] M.-F. Bidaut-Véron, "Local and global behavior of solutions of quasilinear equations of Emden–Fowler type", *Arch. Rational Mech. Anal.* **107**:4 (1989), 293–324. MR Zbl

[Bidaut-Véron and Pohozaev 2001] M.-F. Bidaut-Véron and S. Pohozaev, "Nonexistence results and estimates for some nonlinear elliptic problems", *J. Anal. Math.* **84** (2001), 1–49. MR Zbl

[Caristi et al. 2008] G. Caristi, L. D'Ambrosio, and E. Mitidieri, "Liouville theorems for some nonlinear inequalities", *Tr. Mat. Inst. Steklova* **260**:1 (2008), 97–118. In Russian; translated in *Proc. Steklov Inst. Math.* **260**:1 (2008), 90–111. MR Zbl

[Caristi et al. 2009] G. Caristi, E. Mitidieri, and S. I. Pohozaev, "Liouville theorems for quasilinear elliptic inequalities", *Dokl. Acad. Nauk.* **424**:6 (2009), 741–747. In Russian; translation in *Dokl. Math.* **79**:1 (2009) 118–124. MR Zbl

[Fujita 1966] H. Fujita, "On the blowing up of solutions of the Cauchy problem for $u_t = \Delta u + u^{1+\alpha}$", *J. Fac. Sci. Univ. Tokyo Sect. I* **13** (1966), 109–124. MR Zbl

[Gidas and Spruck 1981] B. Gidas and J. Spruck, "Global and local behavior of positive solutions of nonlinear elliptic equations", *Comm. Pure Appl. Math.* **34**:4 (1981), 525–598. MR Zbl

[Hayakawa 1973] K. Hayakawa, "On nonexistence of global solutions of some semilinear parabolic differential equations", *Proc. Japan Acad.* **49** (1973), 503–505. MR Zbl

[Kobayashi et al. 1977] K. Kobayashi, T. Sirao, and H. Tanaka, "On the growing up problem for semilinear heat equations", *J. Math. Soc. Japan* **29**:3 (1977), 407–424. MR

[Laptev 2003] G. G. Laptev, "Nonexistence results for higher-order evolution partial differential inequalities", *Proc. Amer. Math. Soc.* **131**:2 (2003), 415–423. MR Zbl

[Mitidieri and Pohozaev 1998] E. Mitidieri and S. I. Pohozaev, "Absence of global positive solutions of quasilinear elliptic inequalities", *Dokl. Akad. Nauk* **359**:4 (1998), 456–460. In Russian; translated in *Dokl. Math.* **57**:2 (1998), 250–253. MR Zbl

[Mitidieri and Pohozaev 2001] E. Mitidieri and S. I. Pohozaev, "A priori estimates and the absence of solutions of nonlinear partial differential equations and inequalities", *Tr. Mat. Inst. Steklova* **234** (2001), 3–383. In Russian; translation in *Proc. Steklov Inst. Math.* **234**:3 (2001), 1–362. MR Zbl

[Ni and Serrin 1986a] W.-M. Ni and J. Serrin, "Existence and non-existence theorems for ground states of quasilinear partial differential equations: the anomalous case", *Atti Conv. Lincei* **77** (1986), 231–257.

[Ni and Serrin 1986b] W.-M. Ni and J. Serrin, "Nonexistence theorems for singular solutions of quasilinear partial differential equations", *Comm. Pure Appl. Math.* **39**:3 (1986), 379–399. MR Zbl

[Zhang 1998] Q. S. Zhang, "Blow up and global existence of solutions to an inhomogeneous parabolic system", *J. Differential Equations* **147**:1 (1998), 155–183. MR Zbl

[Zhang 1999] Q. S. Zhang, "Blow-up results for nonlinear parabolic equations on manifolds", *Duke Math. J.* **97**:3 (1999), 515–539. MR Zbl

[Zhang 2001] Q. S. Zhang, "A general blow-up result on nonlinear boundary-value problems on exterior domains", *Proc. Roy. Soc. Edinburgh Sect. A* **131**:2 (2001), 451–475. MR Zbl

Received July 3, 2016. Revised July 28, 2017.

YUHUA SUN
SCHOOL OF MATHEMATICAL SCIENCES AND LPMC
NANKAI UNIVERSITY
TIANJIN
CHINA
sunyuhua@nankai.edu.cn

Guidelines for Authors

Authors may submit articles at msp.org/pjm/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use LaTeX, but papers in other varieties of TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a website in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PACIFIC JOURNAL OF MATHEMATICS

Volume 293 No. 1 March 2018