Pacific Journal of Mathematics

Volume 303 No. 1 November 2019

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 2019 is US $490/year for the electronic version, and $665/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 1945-5844 electronic, 0030-8730 printed) 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. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.1

CONTRASTING VARIOUS NOTIONS OF CONVERGENCE IN GEOMETRIC ANALYSIS

BRIAN ALLEN AND CHRISTINA SORMANI

We explore the distinctions between L^p convergence of metric tensors on a fixed Riemannian manifold versus Gromov–Hausdorff, uniform, and intrinsic flat convergence of the corresponding sequence of metric spaces. We provide a number of examples which demonstrate these notions of convergence do not agree even for two dimensional warped product manifolds with warping functions converging in the L^p sense. We then prove a theorem which requires L^p bounds from above and C^0 bounds from below on the warping functions to obtain enough control for all these limits to agree.

1. Introduction

When mathematicians have studied sequences of Riemannian manifolds arising naturally in questions of almost rigidity or when searching for solutions to geometric partial differential equations, they have obtained bounds on the metric tensors of these Riemannian manifolds. When the bounds they obtained on (M^n, g_j) guaranteed a subsequence, $g_j \to g_\infty$ converging in the C^0 sense or stronger, then the Riemannian manifolds, (M, g_j), viewed as metric spaces, (M, d_j), converge uniformly to (M, d_∞) where d_∞ is defined as the infimum of the lengths of curves between points measured using g_∞. After observing this, Gromov [1981] introduced the Gromov–Hausdorff distance between metric spaces, proving that uniform convergence implies Gromov–Hausdorff convergence of metric spaces. The advantage

C. Sormani was partially supported by NSF DMS 1612049.
MSC2010: 53C23.
Keywords: Gromov–Hausdorff convergence, Sormani–Wenger intrinsic flat convergence, convergence of Riemannian manifolds, warped products.

of Gromov–Hausdorff convergence is that one may allow the spaces themselves to change (M_j, d_j) and one may obtain a limit metric space which is not even a manifold. Gromov [1981] proved that if (M_j, g_j) have uniform lower bounds on Ricci curvature and uniform upper bounds on diameter then a subsequence converges in the Gromov–Hausdorff sense to a metric space, and since then many people have analyzed the properties of these limit spaces.

More recently Sormani and Wenger [2011] introduced the intrinsic flat distance between oriented Riemannian manifolds which need not be diffeomorphic. Roughly the intrinsic flat distance is measuring a filling volume between two manifolds. A standard sphere and a sphere with a thin deep well are very close in the intrinsic flat sense based on the filling volume of the well, while they are far apart in the Gromov–Hausdorff distance based on the depth of the well. As soon as this notion was introduced people began asking whether L^p convergence of the metric tensors might in some way be related to intrinsic flat convergence of the metric spaces. After all, a uniform L^n bound on metric tensors implies a uniform upper bound on volume. Wenger [2011] proved that as long as a sequence of oriented Riemannian manifolds has a uniform upper bound on volume and on diameter it has a subsequence converging in the intrinsic flat sense. However it is not known whether the limit space is in anyway related to (M, g_∞) even when g_∞ was smooth. In joint work with Lakzian [Lakzian and Sormani 2013], and work of Lakzian alone [2016] it was shown that even when $g_j \to g_\infty$ smoothly away from a singular set, the Gromov–Hausdorff and intrinsic flat limits need not be closely related to (M, g_∞) unless one controls volumes, areas, and distances near the singular set.

In this paper we provide a number of examples demonstrating that when metric tensors g_j converge in the L^p sense to a metric tensor g_∞, then uniform, intrinsic flat and Gromov–Hausdorff limits need not converge to a metric space which is defined by g_∞ using the infimum of lengths over all curves. Our examples include very simple two dimensional warped product Riemannian manifolds whose metric tensors are of the form $dr^2 + f_j(r)^2\, d\theta^2$.

In Example 3.4 we find a sequence of warping functions $f_j(r)$ which converge in the L^p sense to a constant function, f_∞, but the uniform, Gromov–Hausdorff, and intrinsic flat limit of the sequence is not even a Riemannian manifold. In this example the $f_j \le f_\infty$ but have an increasingly narrow dip downward about $r = 0$ so we say the sequence of manifolds is "cinched" at 0. This is an example with smooth convergence away from a singular set that was not seen in [Lakzian and Sormani 2013]. The limit metric space is described in detail within the example and a proof is given afterwards. In Example 3.5 the $f_j \le f_\infty$ and L^p converge to f_∞ again, but the cinch moves around so that the f_j do not converge pointwise almost everywhere. This example has no uniform, Gromov–Hausdorff, or intrinsic flat limit unless one takes a subsequence where the cinch's location converges.

In Examples 3.7–3.9 we also consider warping functions, f_j, that L^p converge to a constant function, f_∞, but now $f_j \geq f_\infty$. In Example 3.7 we have a single increasingly narrow peak about $r = 0$. We say there is a "ridge" at 0. This is another example with smooth convergence away from a singular set that was not studied in [Lakzian and Sormani 2013]. We observe how the shortest paths between points on the ridge, do not lie on the ridge in Lemma 3.6. In Example 3.8 we have a sequence of manifolds with moving ridges, so there is no pointwise convergence almost everywhere. In Example 3.9 we have increasingly many increasingly dense ridges. In all three of these examples we prove uniform convergence of the distances, d_j, to d_∞ of the isometric product Riemannian manifold with metric tensor $g_\infty = dr^2 + f_j(r)^2\, d\theta^2$. We obtain intrinsic flat and Gromov–Hausdorff convergence to this limit as well.

In Example 3.12 we have $f_j \geq f_\infty$ with f_∞ constant and $f_j = f_\infty$ on an increasingly dense set. However, now our f_j do not converge in L^p to f_∞. For the particular sequence we chose, we obtain uniform, intrinsic flat and Gromov–Hausdorff convergence to a non-Riemannian Finsler manifold we call a minimized R-stretched Euclidean taxi metric space. This metric is defined as an infimum over an interpolation between a Euclidean metric stretched by R in one direction and a taxi metric. Our example demonstrates that the L^p convergence was crucial in the prior examples. As discussed in Remark 3.13, this example shows the necessity of scalar curvature bounds in the statement of the scalar compactness conjecture of Gromov and Sormani (see [Gromov 2018]) to conclude that the limit has Euclidean tangent cones almost everywhere. This conjecture was recently verified in the rotationally symmetric case by Park, Tian, and Wang [Park et al. 2018].

We then prove the following general theorem concerning warped product manifolds $M^n = [r_0, r_1] \times_f \Sigma$ where Σ is an $n-1$ dimensional manifold including also M without boundary that have f periodic with period $r_1 - r_0$ as in (1)):

Theorem 1.1. *Assume the warping factors, $f_j \in C^0(r_0, r_1)$, satisfy the following*:

$$0 < f_\infty(r) - \frac{1}{j} \leq f_j(r) \leq K < \infty \qquad \text{and} \qquad f_j(r) \to f_\infty(r) > 0 \quad \text{in } L^2,$$

where $f_\infty \in C^0(r_0, r_1)$.

Then we have GH and $\mathcal{F}$ convergence of the warped product manifolds,

$$M_j = [r_0, r_1] \times_{f_j} \Sigma \to M_\infty = [r_0, r_1] \times_{f_\infty} \Sigma,$$
$$N_j = \mathbb{S}^1 \times_{f_j} \Sigma \to N_\infty = \mathbb{S}^1 \times_{f_\infty} \Sigma,$$

and uniform convergence of their distance functions, $d_j \to d_\infty$.

Remark 1.2. In our theorem we assume L^2 convergence but since we are assuming that the f_j are uniformly bounded this is equivalent to L^p, $p \in [1, \infty)$ convergence.

The proof of this theorem and indeed the proof of all the examples relies on a theorem of the second author with Huang and Lee in the appendix of [Huang et al. 2017] which is reviewed in the background section of this paper. The theorem in [Huang et al. 2017] states that if one has uniform upper and lower bounds on the d_j, a subsequence of the Riemannian manifolds converges in the uniform, Gromov–Hausdorff, and intrinsic flat convergence sense to some common limit space. Thus we need only prove pointwise convergence of the original sequence of d_j to our proposed d_∞. The method applied to control d_j is different in each proof in this paper. For the theorem, we apply the C^0 lower bound to bound d_j from below and the L^p upper bound is all that is needed to bound d_j from above pointwise. Note that the hypothesis of the theorem immediately implies a uniform upper bound on diameter (Lemma 4.2). We end the paper with Theorem 5.1 concerning warped product manifolds where the warping function depends on two variables.

Applications of these theorems will appear in a paper by the first author with Hernandez-Vazquez, Parise, Payne, and Wang [Allen et al. 2019] on a conjecture of Gromov concerning the almost rigidity of the scalar torus theorem. The first author hopes to apply the techniques developed here in combination with his prior work in [Allen 2018a; 2018b] to prove a special case of Lee and Sormani's conjecture [2014] on almost rigidity of the positive mass theorem. Additional applications to conjectures involving scalar curvature that were raised by the second author at the Fields Institute and described in [Sormani 2017] will be explored with other teams of students and postdocs in the near future. Anyone interested in joining one of these teams should contact the second author.

2. Review

In this subsection we review what we mean by a warped product space even with a noncontinuous warping function and what one needs to know about Gromov–Hausdorff and intrinsic flat convergence to prove all examples and theorems in this paper. The reader does not need any prior knowledge of these two notions of convergence. Readers who are experts in these notions of convergence are recommended to read just the first and last subsections of this review section of the paper, particularly Theorem 2.4 which combines results of Gromov [1981] and the second author with Huang and Lee [Huang et al. 2017]. All examples and theorems in this paper apply that theorem to prove convergence.

2A. ***Warped product spaces.*** Let (Σ^{n-1}, σ) be a compact Riemannian manifold and

$$f : [r_1, r_2] \to \mathbb{R}^+$$

and define the warped product manifolds

$$M = [r_1, r_2] \times_f \Sigma \quad \text{and} \quad N = \mathbb{S}^1 \times_f \Sigma \tag{1}$$

with warped product metrics defined by

$$g = dr^2 + f^2(r)\sigma, \tag{2}$$

where either $r \in [r_1, r_2]$ or $r \in \mathbb{S}^1$. On such a manifold we define lengths of curves to be

$$L_g(C) = \int_0^1 g(C'(t), C'(t))^{1/2}\, dt = \int_0^1 \sqrt{|r'(t)|^2 + |f(r(t))|^2|\theta'(t)|^2}\, dt \tag{3}$$

which is well defined even when f is only L^1. We then define distances $d_g^M(p, q)$ and $d_g^N(p, q)$ on M and N respectively as

$$d_g(p, q) = \inf\{L_g(C) : C(0) = p,\ C(1) = q\} \tag{4}$$

where the value is different on M and N because the selection of curves between points within these two spaces are different.

Remark 2.1. Note that we do not need f to be smooth or even continuous to define a warped product metric space. As long as the function is bounded above, we can define lengths using (3). Following the text of Burago, Burago, and Ivanov [Burago et al. 2001], the distance d defined by (4) is symmetric and satisfies the triangle inequality. It is positive definite as long as f is bounded below by a positive number. Such a metric space is then compact and there are geodesics whose lengths achieve the infimum in (4). Even more general warped products of metric spaces are explored by Alexander and Bishop [2004].

Remark 2.2. Throughout this paper we will assume that our warping function f is continuous. Annegret Burtscher has proven that if a Riemannian manifold has a continuous metric tensor then the distance between points is achieved by an absolutely continuous curve (See Proposition 4.1 and Theorem 4.11 in [Burtscher 2015]). This is achieved by showing that the length of absolutely continuous curves defined by (3) is equivalent to the induced length (see [loc. cit., Definition 2.1]) defined by d_g in [loc. cit., Theorem 4.11]. This will be important for us because we will repeatedly use the fact that the distance between points of M can be achieved by an absolutely continuous curve $C(t)$ and hence we can reparametrize $C(t)$ so that $|C'(t)|_g = 1$ almost everywhere.

For warped products we can show that L^2 convergence of metrics $g_j \to g_\infty$ is equivalent to L^2 convergence of the warping functions $f_j \to f_\infty$. For this we fix the background metric $\delta = dr^2 + \sigma$ and an orthonormal basis for this metric

$\{\partial_r, \partial_{\theta_1}, \ldots, \partial_{\theta_n}\}$ and compute

$$\int_M |g_j - g_\infty|_\delta^2 \, dm = \int_M \sum_{i=1}^{n} |f_j - f_\infty|^2 \sigma(\partial_{\theta_i}, \partial_{\theta_i}) \, dm$$
$$= n \int_{r_1}^{r_2} \int_\Sigma |f_j - f_\infty|^2 \, d\mu \, dr = n|\Sigma| \int_{r_1}^{r_2} |f_j - f_\infty|^2 \, dr,$$

where dm is the measure on M induced by δ, $d\mu$ is the measure on Σ from σ and $|\Sigma|$ is n-dimensional volume of Σ. This shows that we can just work with L^2 convergence of the warping functions for the sake of this paper.

2B. ***Gromov–Hausdorff convergence.*** Gromov–Hausdorff convergence was introduced by Gromov [1981]. See also the text of Burago–Burago–Ivanov [Burago et al. 2001]. It measures a distance between metric spaces. It is an intrinsic version of the Hausdorff distance between sets in a common metric space Z:

$$d_H^Z(A_1, A_2) = \inf\{r : A_1 \subset T_r(A_2) \text{ and } A_2 \subset T_r(A_1)\},$$

where $T_r(A) = \{x \in Z : \exists a \in A \text{ s.t. } d_Z(x, a) < r\}$. Since an arbitrary given pair of compact metric spaces, (X_i, d_i), might not lie in the same compact metric space, we use distance preserving maps:

$$\varphi_i : X_i \to Z \quad \text{such that} \quad d_Z(\varphi_i(p), \varphi_i(q)) = d_i(p, q) \quad \text{for all } p, q \in X_i$$

to map them into a common compact metric space, Z.

The Gromov–Hausdorff distance between two compact metric spaces, (X_i, d_i), is then defined to be

$$d_{\mathrm{GH}}((X_1, d_1), (X_2, d_2)) = \inf\{d_H^Z(\varphi_1(X_1), \varphi_2(X_2)) : \varphi_i : X_i \to Z\},$$

where the infimum is taken over all compact metric spaces Z and all distance preserving maps, $\varphi_i : X_i \to Z$.

2C. ***Warped products as integral current spaces.*** Intrinsic flat convergence is defined for sequences of integral current spaces in [Sormani and Wenger 2011]. An integral current space is a metric space, (X, d), endowed with a current structure, T, where T is defined by a collection of bi-Lipschitz charts with weights. If we start with an oriented smooth Riemannian manifold, M, then (X, d) is the standard metric space defined by M using lengths of curves as in (3) and T is defined by the orientation of M,

$$T(f, \pi_1, \ldots, \pi_m) = \int_M f \, d\pi_1 \wedge \cdots \wedge d\pi_m. \tag{5}$$

Here we are considering warped product spaces, M and N, as in (1) allowing our function, $f : [r_1, r_2] \to \mathbb{R}^+$, to simply have a maximum and a positive minimum and

do not require it to be smooth. In order to confirm that we still may use (5) to define the integral current structure on our space, we need only verify that our standard oriented charts on the isometric product manifold are bi-Lipschitz to the metric d we obtain as in (3)–(4). This is confirmed by showing the identity map between the isometric product manifold, $M_1 = [r_1, r_2] \times_1 \Sigma$, and our warped product space, $M = [r_1, r_2] \times_f \Sigma$, is bi-Lipschitz:

Lemma 2.3. *Suppose the warping function is bounded*

$$f(r) \in [a, b] \quad \textit{for all } r \in [r_1, r_2],$$

then the identity map

$$F : M_1 = [r_1, r_2] \times_1 \Sigma \to M = [r_1, r_2] \times_f \Sigma$$

is bi-Lipschitz:

$$0 < \min\{a, 1\} \leq \frac{d_M(F(p), F(q))}{d_{M_1}(p, q)} \leq (\max\{1, b\}).$$

Proof. This can be seen by observing that

$$\begin{aligned} L_g(C) &= \int_0^1 \sqrt{|r'(t)|^2 + |f(r(t))|^2 |\theta'(t)|^2}\, dt \\ &\leq (\max\{1, b\}) \int_0^1 \sqrt{|r'(t)|^2 + |\theta'(t)|^2}\, dt \\ &\leq (\max\{1, b\}) L_{g_1}(C). \end{aligned}$$

Thus

$$d_M(F(p), F(q)) \leq (\max\{1, b\})\, d_{M_1}(p, q).$$

For the other direction we have

$$\begin{aligned} L_{g_1}(C) &= \int_0^1 \sqrt{|r'(t)|^2 + |\theta'(t)|^2}\, dt \\ &\leq (\min\{a, 1\})^{-1} \int_0^1 \sqrt{|r'(t)|^2 + |f(r(t))|^2 |\theta'(t)|^2}\, dt \\ &\leq (\min\{a, 1\})^{-1} L_g(C). \end{aligned}$$

Thus

$$d_{M_1}(p, q) \leq (\min\{a, 1\})^{-1} d_M(F(p), F(q)). \qquad \square$$

2D. ***Key theorem we apply to prove GH and $\mathcal{F}$ convergence.*** The following theorem was proven by the second author jointly with Huang and Lee in [Huang et al. 2017] building upon earlier work of Gromov [1981]. This theorem allows us to prove GH and intrinsic flat convergence using only information about the sequence of distance functions. Note that it is a compactness theorem, providing the existence of a converging subsequence once one simply has uniform bi-Lipschitz control

on the metrics. The convergence is not bi-Lipschitz convergence but instead it is uniform convergence of the distance functions and also GH and $\mathcal{F}$ convergence of the spaces.

Theorem 2.4. *Fix a precompact n-dimensional integral current space (X, d_0, T) without boundary* (e.g., $\partial T = 0$) *and fix $\lambda > 0$. Suppose that d_j are metrics on X such that*

$$\lambda \geq \frac{d_j(p,q)}{d_0(p,q)} \geq \frac{1}{\lambda}. \tag{6}$$

Then there exists a subsequence, also denoted d_j, and a length metric d_∞ satisfying (6) *such that d_j converges uniformly to d_∞:*

$$\epsilon_j = \sup\big\{|d_j(p,q) - d_\infty(p,q)| : p, q \in X\big\} \to 0.$$

Furthermore

$$\lim_{j\to\infty} d_{\mathrm{GH}}\big((X, d_j), (X, d_\infty)\big) = 0$$

and

$$\lim_{j\to\infty} d_{\mathcal{F}}\big((X, d_j, T), (X, d_\infty, T)\big) = 0.$$

In particular, (X, d_∞, T) is an integral current space and $\mathrm{set}(T) = X$ *so there are no disappearing sequences of points $x_j \in (X, d_j)$.*

In fact we have

$$d_{\mathrm{GH}}\big((X, d_j), (X, d_\infty)\big) \leq 2\epsilon_j$$

and

$$d_{\mathcal{F}}\big((X, d_j, T), (X, d_\infty, T)\big) \leq 2^{(n+1)/2}\lambda^{n+1}2\epsilon_j \boldsymbol{M}_{(X,d_0)}(T).$$

Remark 2.5. In order to apply this theorem we will use the following method repeatedly. We will demonstrate that a sequence has pointed convergence of the distance functions and also satisfies the bi-Lipschitz bound in (6). Then by this theorem there is a converging subsequence. However by the pointed convergence we will see that all the subsequences must in fact converge to the same limit space. Thus we obtain $\mathcal{F}$ and GH convergence of the original sequence.

3. Examples

In this section we present our examples. Each example contains a sequence of smooth warped product manifolds which converge in various ways to warped product metric spaces. We first study distances on warped product spaces with deep valleys. We apply this to present our cinched warped product example. We then observe what happens to distances on warped product spaces with peaks.

3A. ***Distances on warped products with valleys.*** First let us develop the intuitive picture first. Consider a warped product manifold $[-\pi, \pi] \times_g \mathbb{S}^1$ as in Figure 1 with a warping function

$$f_j(r) = \begin{cases} 1, & r \in [-\pi, -1/j], \\ h(jr), & r \in [-1/j, 1/j], \\ 1, & r \in [1/j, \pi], \end{cases}$$

where h is a smooth even function defining a valley with $h(-1) = 1$ with $h'(-1) = 0$, decreasing to $h(0) = h_0 \in (0, 1]$ and then increasing back up to $h(1) = 1$, $h'(1) = 0$. Keep in mind that the distance between the level sets, $r^{-1}(a)$ and $r^{-1}(b)$ is $|a - b|$ and so we have evenly spaced levels drawn in the figure.

A minimizing geodesic, draw in red in Figure 1, will proceed diagonally towards the valley, climb down into the valley, run along the valley, then climb out and proceed diagonally away from the valley. The climbing parts are very short if the change in r is small (which is true for large j). Since it is more efficient to travel around inside the valley (for the change in θ), it is more efficient to travel almost directly to the valley as in the geodesic in the figure. Observe that the length of this geodesic is bounded above by the length of a curve which goes directly to the valley and straight down, then turns a right angle to stay along the bottom of the valley, and then makes a right angle to climb out and move directly to the end point. Thus

$$d((-r, \theta_1), (r, \theta_2)) \leq |-r - 0| + f(0)\, d_{\mathbb{S}^1}(\theta_1, \theta_2) + |0 - r|.$$

In the following lemmas we use this same basic idea to bound distances in warped products with a wide variety of warping functions.

Lemma 3.1. *Given a warped product space* M (or respectively N) *defined as in* (1), *suppose* $f(r) \geq f(r_0)$ *for all* $r \in [r_1, r_2]$ (or respectively $r \in \mathbb{S}^1$). *If* $x_1, x_2 \in r^{-1}(r_0)$ *then*

$$d_g(x_1, x_2) = f(r_0)\, d_\sigma(\theta_2, \theta_1).$$

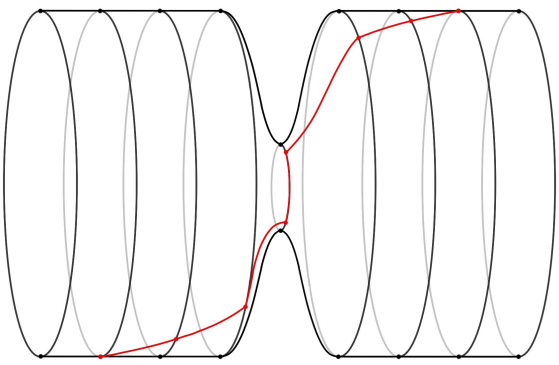


Figure 1. The geodesic will cut across the valley.

Proof. Let $C(t) = (r(t), \theta(t))$ be any curve joining $x_1 = (r_0, \theta_1)$ to $x_2 = (r_0, \theta_2)$. Then

$$\begin{aligned} L(C[0,1]) &= \int_0^1 \sqrt{|r'(t)|^2 + |f(r(t))|^2 |\theta'(t)|^2}\, dt \\ &\geq \int_0^1 \sqrt{|0|^2 + |f(r_0)|^2 |\theta'(t)|^2}\, dt \\ &= f(r_0) \int_0^1 |\theta'(t)|\, dt \\ &= f(r_0)\, L_\Sigma(\theta[0,1]) \\ &\geq f(r_0)\, d_\sigma(\theta_2, \theta_1). \end{aligned}$$

However if we take the curve $C(t) = (r_0, \theta(t))$, where $\theta(t)$ is a minimizing geodesic in Σ from θ_1 to θ_2, we have equality everywhere above. So the infimum over all lengths is achieved:

$$d_g(x_1, x_2) = \inf_C L(C[0,1]) = f(r_0) d_\sigma(\theta_2, \theta_1). \qquad \square$$

Lemma 3.2. *Given a warped product space M defined as in* (1) *and a pair of points $x_1 = (r_1, \theta_1)$ and $x_2 = (r_2, \theta_2)$ with $r_1 < r_2$ then the distance between those points is bounded by*

$$d^M_{g_j}(x_1, x_2) \leq |r_2 - r_1| + D_j(r_1, r_2) d_\sigma(\theta_2, \theta_1),$$

where

$$D_j(r_1, r_2) = \min_{r \in [r_1, r_2]} f_j(r)$$

and d_σ is the distance on (Σ, σ).

Proof. Let $\hat{r}_j \in (r_1, r_2)$ be chosen so that $f_j(\hat{r}_j) = D_j(r_1, r_2)$. Construct the following curve between the points $x_1, x_2 \in M_j$, where $\alpha \subset \Sigma$ is a geodesic with respect to (Σ, σ), $\alpha(0) = \theta_1$ and $\alpha(1) = \theta_2$,

$$C_j(t) = \begin{cases} (r_1 + 3(\hat{r}_j - r_1)t, \theta_1), & t \in \left[0, \frac{1}{3}\right], \\ (\hat{r}_j, \alpha(3t-1)), & t \in \left[\frac{1}{3}, \frac{2}{3}\right], \\ \left(\hat{r}_j + 3(r_2 - \hat{r})\left(t - \frac{2}{3}\right), \theta_2\right), & t \in \left[\frac{2}{3}, 1\right], \end{cases}$$

and then

$$d^M_{g_j}(x_1, x_2) \leq L_j(C_j) = |r_2 - \hat{r}_j| + f_j(\hat{r}_j) d_\sigma(\theta_2, \theta_1) + |\hat{r}_j - r_1|. \qquad \square$$

Almost the same proof can be applied to show the following lemma.

Lemma 3.3. *Given a warped product space N defined as in* (1) *and a pair of points $x_1 = (r_1, \theta_1)$ and $x_2 = (r_2, \theta_2)$ then the distance between those points is bounded by*

$$d^M_{g_j}(x_1, x_2) \leq d_{\mathbb{S}^1}(r_1, r_2) + D_j(r_1, r_2) d_\sigma(\theta_2, \theta_1),$$

where

$$D_j(r_1, r_2) = \min_{r \in \mathrm{arc}(r_1, r_2)} f_j(r),$$

where $\mathrm{arc}(r_1, r_2)$ *is the minor arc between* r_1 *and* r_2 *in* $\mathbb{S}^1$ *and where* d_σ *is the distance on* (Σ, σ).

3B. ***Cinched spaces.*** Here we see examples of spaces whose warping functions converge in the L^p sense but the GH and SWIF limits do not agree with the L^p limit due to the existence of deep canyons or cinching. See Figure 1 and now imagine that the valley remains equally as deep but becomes very narrow.

Example 3.4. Consider the sequence of smooth functions $f_j(r) : [-\pi, \pi] \to [1, 2]$

$$f_j(r) = \begin{cases} 1, & r \in [-\pi, -1/j], \\ h(jr), & r \in [-1/j, 1/j], \\ 1, & r \in [1/j, \pi], \end{cases}$$

where h is a smooth even function such that $h(-1) = 1$ with $h'(-1) = 0$, decreasing to $h(0) = h_0 \in (0, 1]$ and then increasing back up to $h(1) = 1,\ h'(1) = 0$. Note that this defines a sequence of smooth Riemannian metrics, g_j, as in (2), with distances, d_j, as in (4) on the manifolds,

$$M_j = [-\pi, \pi] \times_{f_j} \Sigma \quad \text{or} \quad N_j = \mathbb{S}^1 \times_{f_j} \Sigma$$

for any fixed Riemannian manifold Σ. Consider also M_∞ and N_∞ defined as above with $f_\infty(r) = 1$ for all r.

Despite the fact that

$$f_j \to f_\infty \quad \text{in } L^p,$$

we do not have M_j converging to M_∞ nor N_j to N_∞ in the GH or $\mathcal{F}$ sense. In fact

$$M_j \xrightarrow{\mathrm{GH}} M_0 \quad \text{and} \quad M_j \xrightarrow{\mathcal{F}} M_0$$

and

$$N_j \xrightarrow{\mathrm{GH}} N_0 \quad \text{and} \quad N_j \xrightarrow{\mathcal{F}} N_0,$$

where M_0 and N_0 are warped metric spaces defined as in (1) with warping factor

$$f_0(r) = \begin{cases} 1, & r \in [-\pi, 0), \\ h_0, & r = 0, \\ 1, & r \in (0, \pi]. \end{cases}$$

Proof. First we verify our claim about L^p convergence

$$\left(\int_{-\pi}^{\pi} |f_j - 1|^p \, dr \right)^{1/p} = \left(\int_{-1/j}^{1/j} |h_j - 1|^p \, dr \right)^{1/p} \leq \left(\frac{2}{j} \right)^{1/p} \to 0,$$

where we use the fact that $|h_j - 1|^p \leq 1$ by construction.

Let us consider (M_j, d_j). Since we have

$$0 < h_0 \leq f_j(r) \leq f_0(r) \leq f_\infty(r) = 1$$

then

$$(h_0)^2 \, g_\infty \leq g_j \leq g_0 \leq g_\infty$$

and

$$h_0 \, d_\infty(x_1, x_2) \leq d_j(x_1, x_2) \leq d_0(x_1, x_2) \leq d_\infty(x_1, x_2).$$

Using d_∞ as our background metric we can apply the theorem in the appendix of [Huang et al. 2017] to see that a subsequence of the d_j converges uniformly to some limit, d, such that

$$h_0 \, d_\infty(x_1, x_2) \leq d(x_1, x_2) \leq d_0(x_1, x_2) \leq d_\infty(x_1, x_2). \tag{7}$$

In addition the subsequences converge in the Gromov–Hausdorff and intrinsic flat sense:

$$(M_j, d_j) \xrightarrow{\text{GH}} (M, d) \quad \text{and} \quad (M_j, d_j, T) \xrightarrow{\mathcal{F}} (M, d, T).$$

We need only prove $d = d_0$ for then no subsequence was necessary and we have proven our example.

Consider $x_1, x_2 \in M$ such that

$$d(x_1, x_2) < \min\{d(x_1, p) + d(p, x_2) : p \in r^{-1}(0)\}.$$

So there exists $\delta > 0$ depending on these two points such that

$$d(x_1, x_2) + \delta \leq \min\{d(x_1, p) + d(p, x_2) : p \in r^{-1}(0)\}.$$

Then for N sufficiently large, and all $j \geq N$ (in our subsequence) we have

$$d_j(x_1, x_2) + \delta/2 \leq \min\{d_j(x_1, p) + d_j(p, x_2) : p \in r^{-1}(0)\}.$$

Thus the L_{g_j}-shortest curve, γ_j, between x_1 and x_2 avoids $r^{-1}(-\delta/4, \delta/4)$. Here we have $g_j = g_0 = g_\infty$ so its length is the same with respect to all three metrics:

$$L_{g_j}(\gamma_j) = L_{g_0}(\gamma_j) = L_{g_\infty}(\gamma_j).$$

So

$$d_j(x_1, x_2) \geq d_0(x_1, x_2)$$

and taking the limit we have

$$d(x_1, x_2) \geq d_0(x_1, x_2)$$

and combining this with (7) we have

$$d(x_1, x_2) = d_0(x_1, x_2).$$

In fact for any L_d-shortest curve γ,

$$\gamma([t_1, t_2]) \cap r^{-1}(0) = \varnothing \implies d(\gamma(t_1), \gamma(t_2)) = d_0(\gamma(t_1), \gamma(t_2)). \tag{8}$$

We need only confirm that $d(x_1, x_2) = d_0(x_1, x_2)$ for $x_1, x_2 \in M$ such that

$$d(x_1, x_2) = \min\{d(x_1, p) + d(p, x_2) : p \in r^{-1}(0)\}.$$

Taking the L_d-shortest curve γ between x_1 and x_2, we know that $s_1 \le s_2$

$$s_1 = \inf\{t : \gamma(t) \in r^{-1}(0)\} \quad \text{and} \quad s_2 = \sup\{t : \gamma(t) \in r^{-1}(0)\}.$$

We have

$$d(x_1, x_2) = L_d(\gamma) = d(\gamma(0), \gamma(s_1)) + d(\gamma(s_1), \gamma(s_2)) + d(\gamma(s_2), \gamma(1))$$

By (8) if $s_1 > 0$ then for all $\delta > 0$ we have

$$d(\gamma(0), \gamma(s_1 - \delta)) = d_0(\gamma(0), \gamma(s_1 - \delta))$$

so

$$d(\gamma(0), \gamma(s_1)) = d_0(\gamma(0), \gamma(s_1)).$$

Similarly

$$d(\gamma(s_2), \gamma(1)) = d_0(\gamma(s_2), \gamma(1)).$$

Thus we need only confirm that $d(x_1, x_2) = d_0(x_1, x_2)$ for $x_1, x_2 \in r^{-1}(0)$. This easily follows by applying Lemma 3.1 to both f_j and f_0 since both functions have minimum $= h_0$ at $r = 0$:

$$d(x_1, x_2) = \lim_{j\to\infty} d_j(x_1, x_2) = h_0 d_\sigma(\theta_1, \theta_2) = d_0(x_1, x_2).$$

To prove the case where we have a warped product of the form N as in (1) the proof is almost the same. □

3C. ***Moving cinches.*** Here we explore what happens when the warping functions converge in L^p but not pointwise almost everywhere.

Example 3.5. We first construct a classical sequence of smooth functions $f_j : [-\pi, \pi] \to (0, 1]$ which converge L^p to $f_\infty = 1$ but do not converge pointwise almost everywhere without taking a subsequence. Let

$$f_j(r) = \begin{cases} h((r - t_j)/\delta_j), & r \in [t_j - \delta_j, t_j + \delta_j], \\ 1, & \text{elsewhere}, \end{cases}$$

where h is a smooth even function as in Example 3.4 such that $h(-1) = 1$ with $h'(-1) = 0$, decreasing to $h(0) = h_0 \in (0, 1]$ and then increasing back up to $h(1) = 1$, $h'(1) = 0$, and where

$$\{t_j : j \in \mathbb{N}\} = \left\{\tfrac{0}{1}, \tfrac{1}{1}, \tfrac{0}{2}, \tfrac{1}{2}, \tfrac{2}{2}, \tfrac{0}{4}, \tfrac{1}{4}, \tfrac{2}{4}, \tfrac{3}{4}, \ldots\right\}$$

and

$$\{\delta_j : j \in \mathbb{N}\} = \left\{\tfrac{1}{1}, \tfrac{1}{1}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \ldots\right\}.$$

Then the cylinders, N_j, defined as in (1) will not converge in the GH or $\mathcal{F}$ sense without taking a subsequence. The tori M_j will converge since each torus in this sequence is isometric to a torus in the sequence of tori in Example 3.4 via an isometry which moves t_j to 0.

Proof. First we check that f_j converges in L^p but not pointwise almost everywhere. To this end we check that

$$\left(\int_{-\pi}^{\pi} |f_j - 1|^p \, dr\right)^{1/p} = \left(\int_{t_j-\delta_j}^{t_j+\delta_j} |h_0 - 1|^p \, dr\right)^{1/p} = (2\delta_j)^{1/p} \to 0$$

since $|h_0 - 1|^p \leq 1$ by construction. Of course we do not find pointwise convergence for any $r \in [0, 1]$ since for every choice of $J > 0$ one can find a $j_1 \geq J$ and a $r \in [-\pi, \pi]$ so that $f_{j_1}(r) = h_0$ and another $j_2 \geq J$ so that $f_{j_2}(r) = 1$.

Now if we take a subsequence where $t_{j_k} = 0$, then exactly as in Example 3.4 we see that N_{j_k} converges in the GH and $\mathcal{F}$ sense to N_0 of that example. On the other hand, if we take a subsequence where $t_{j'_k} = 1$, then imitating the proof in Example 3.4 we see that $N_{j'_k}$ converges in the GH and $\mathcal{F}$ sense to N'_0 which is a warped product whose warping function is 1 everywhere except at $r = 1$ where it is h_0. Thus the original sequence of N_j of this example has no GH nor $\mathcal{F}$ limit. □

3D. ***Avoiding ridges.*** The cinched spaces of Example 3.4 did not converge to their L^p limit because their warping functions, f_j, all had a minimum uniformly below the level of their L^p limit, f_∞. Here we will see there is no corresponding problem when the f_j have a maximum uniformly above the level of their L^p limit.

In the following lemma, we have a ridge as in Figure 2, the minimal geodesic between points, p, q lying on that ridge, will not run along the ridge. In the following we consider f_j with a maximum at r_* and thus there is a ridge along the level set $f_j^{-1}(r_*)$.

Lemma 3.6. *Given $r_*, \hat{r} \in [r_0, r_1]$, the distance between $x_1 = (r_*, \theta_1)$ and $x_2 = (r_*, \theta_2)$ in a warped product space is bounded above by*

$$d(x_1, x_2) \leq 2|\hat{r} - r_*| + f_j(\hat{r})\, d_\sigma(\theta_1, \theta_2).$$

Thus for a fixed $r_* \in [r_0, r_1]$, if there exists an $\hat{r} \in [r_0, r_1]$ such that

$$f_j(\hat{r}) < f_j(r_*) - 2\frac{|\hat{r} - r_*|}{d_\sigma(\theta_1, \theta_2)} \tag{9}$$

then the minimizing geodesic from $x_1 = (r_*, \theta_1)$ to $x_2 = (r_*, \theta_2)$, $\theta_1, \theta_2 \in \Sigma$, $\theta_1 \neq \theta_2$, cannot be a curve with constant r-component, $r(t) = r_*$.

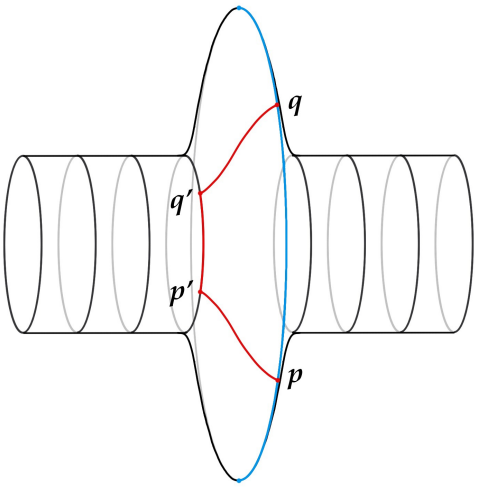


Figure 2. A curve γ from p to q on a ridge, which first cuts down to p' and then runs across to q' before cutting up to q is shorter than curve running along the ridge between p and q if the ridge is narrow enough.

See Figure 2 taking $p = x_1 = (r_*, \theta_1)$ and $q = x_2 = (r_*, \theta_2)$ and $p' = (\hat{r}, \theta_1)$ and $q = x_2 = (\hat{r}, \theta_2)$. So $d(p, q) \leq L(\gamma) = d(p, p') + d(p', q') + d(q', q)$, where $d(p, p') = d(q, q') = |r_* - \hat{r}|$.

Proof. Let $x_1, x_2 \in M_j$ with coordinates $x_1 = (r_*, \theta_1)$ to $x_2 = (r_*, \theta_2)$, $\theta_1, \theta_2 \in \Sigma$, $\theta_1 \neq \theta_2$ so that (9) is satisfied for r_*. Let $\alpha \subset \Sigma$ be a curve between θ_1, θ_2 with length $L_\Sigma(\alpha) = d_\sigma(\theta_1, \theta_2)$ and consider the curve

$$\gamma(t) = \begin{cases} (r_* + 3(\hat{r} - r_*)t, \theta_1), & t \in \left[0, \frac{1}{3}\right], \\ (\hat{r}, \alpha(3t - 1)), & t \in \left[\frac{1}{3}, \frac{2}{3}\right], \\ \left(\hat{r} + 3(r_* - \hat{r})\left(t - \frac{2}{3}\right), \theta_2\right), & t \in \left[\frac{2}{3}, 1\right], \end{cases}$$

as depicted in Figure 2. Then

$$L_j(\gamma) = 2|\hat{r} - r_*| + f_j(\hat{r})\, d_\sigma(\theta_1, \theta_2).$$

So if we consider $\beta(t) = (r_*, \alpha(t))$ and use the assumption (9) then we find that

$$L_j(\gamma) < L_j(\beta)$$

and hence $\beta(t)$ cannot be the minimizing geodesic. □

3E. ***A single ridge disappears.*** Here we see that a sequence of warped product spaces with a consistently high ridge that is increasingly narrow converges in the L^p, pointwise a.e., GH, and $\mathcal{F}$ sense to an isometric product manifold as if the

ridge simply disappears despite the fact that the warping functions do not converge pointwise to the constant function 1. See Figure 2.

Example 3.7. Consider the sequence of functions $f_j(r) : [-\pi, \pi] \to [1, 2]$ with

$$f_j(r) = \begin{cases} 1, & r \in [-\pi, -1/j], \\ h(jr), & r \in [-1/j, 1/j], \\ 1, & r \in [1/j, \pi], \end{cases}$$

where $h = h_{ridge}$ is a smooth even function such that $h(-1) = 1$ with $h'(-1) = 0$, increasing to $h(0) = h_0 \in (1, 2]$ and then decreasing back down to $h(1) = 1$, $h'(1) = 0$. Note that this defines a sequence of smooth Riemannian metrics, g_j, as in (2), with distances, d_j, as in (4) on the manifolds,

$$M_j = [-\pi, \pi] \times_{f_j} \Sigma \quad \text{or} \quad N_j = \mathbb{S}^1 \times_{f_j} \Sigma,$$

for any fixed Riemannian manifold Σ. Consider also M_∞ and N_∞ defined as above with $f_\infty(r) = 1$ for all r. Here we have

$$f_j \to f_\infty = 1 \quad \text{in } L^p \text{ but not pointwise}$$

and yet $M_j \to M_\infty$ and $N_j \to N_\infty$ in both the GH and $\mathcal{F}$ sense.

Proof. First we check that f_j converges in L^p to f_∞. To this end we check that

$$\left(\int_{-\pi}^{\pi} |f_j - f_\infty|^p \, dr \right)^{1/p} = \left(\int_{-1/j}^{1/j} |h(jr) - 1|^p \, dr \right)^{1/p} \leq (2/j)^{1/p} \to 0$$

since $|h_j - 1|^p \leq 1$ by construction. Observe that f_j does not converge pointwise to f_∞ because $f_j(0) = h_0 > 1 = f_\infty(0)$. Let

$$J_\delta = 1/\delta \tag{10}$$

so that $f_j(r) = f_\infty(r)$ on $[0, -1/j] \cup [1/j, 1]$ for all $j \geq J_\delta$.

Next observe that since $2 f_\infty(r) \geq f_j(r) \geq f_\infty(r)$ at all $r \in [-\pi, \pi]$, we have

$$d_\infty(p, q) \leq d_j(p, q) \leq 2 d_\infty(p, q) \quad \text{for all } p, q. \tag{11}$$

Since our limit space, M_∞, is an isometric product space, any pair of points $x_1 = (s_1, \theta_1)$ to $x_2 = (s_2, \theta_2)$ with $s_1 < s_2$ is joined by a smooth L_∞ minimizing geodesic, $C : [0, 1] \to M_\infty$, such that

$$d_\infty(p, q) = L_\infty(C).$$

In fact $C(t) = (r(t), \theta(t))$ where $r : [0, 1] \to [r_1, r_2]$ is strictly increasing from s_1 to s_2, and $\theta : [0, 1] \to \Sigma$ is a geodesic from θ_1 to θ_2 with respect to (Σ, σ). Let $T_\delta \subset [0, 1]$ be defined as the possibly empty interval

$$T_\delta = \{t : r(t) \in [-\delta, \delta]\}.$$

Observe that the length of C restricted to the interval T_δ satisfies

$$L_\infty(C(T_\delta)) \leq 2\delta L_\infty(C) \leq 2\delta d_\infty(x_1, x_2).$$

For $j \geq J_\delta$ as in (10), we have

$$\begin{aligned} d_j(x_1, x_2) \leq L_j(C) &= \int_0^1 g_j(C'(t), C'(t))^{1/2}\, dt \\ &\leq \int_{T_\delta} 2g_\infty(C'(t), C'(t))^{1/2} + \int_{[0,1]\setminus T_\delta} g_\infty(C'(t), C'(t))^{1/2} \\ &\leq 2L_\infty(C(T_\delta)) + L_\infty(C[0,1]) \\ &\leq (1+2\delta)\, d_\infty(x_1, x_2). \end{aligned}$$

Thus for x_1 and x_2 lying on different levels of r we have pointwise convergence $d_j(x_1, x_2) \to d_\infty(x_1, x_2)$.

Taking points that lie on the same level, $x_1 = (s, \theta_1)$ to $x_2 = (s, \theta_2)$, we know that the minimizing geodesic, C, in our isometric product will have the form $C(t) = (s, \theta(t))$. If the points do not lie on the ridge, $s \neq 0$, and so

$$d_j(x_1, x_2) \leq L_j(C) = L_\infty(C) = d_\infty(x_1, x_2) \quad \text{for all } j \geq J_\delta.$$

So again we have pointwise convergence $d_j(x_1, x_2) \to d_\infty(x_1, x_2)$.

If the points both lie on the ridge $x_1 = (0, \theta_1)$ to $x_2 = (0, \theta_2)$ then by Lemma 3.6 we have

$$d_j(x_1, x_2) \leq 1 d_\Sigma(\theta_1, \theta_2) + 2\delta = d_\infty(x_1, x_2) + 2\delta \quad \text{for all } j \geq J_\delta.$$

And again we have pointwise convergence $d_j(x_1, x_2) \to d_\infty(x_1, x_2)$.

By Theorem 2.4 combined with (11) we know a subsequence d_{j_k} converges uniformly to some limit distance. Since we have pointwise convergence to d_∞, we know in fact that the d_j thus converge uniformly to d_∞ without even taking a subsequence. Furthermore we have Gromov–Hausdorff and intrinsic flat convergence.

The proof when we have warped around $\mathbb{S}^1$ to create N_j is very similar. □

3F. ***Moving ridges.*** Here we see a sequence of spaces which have f_j converging to $f_\infty = 1$ in the L^p sense and $f_j \geq 1$. The sequence does not converge pointwise almost everywhere unless one takes a subsequence. Nevertheless by Theorem 1.1 there is a GH and a SWIF limit without taking a subsequence and indeed the limit is the space warped by f_∞.

Example 3.8. We first construct a classical sequence of smooth functions $f_j : [-\pi, \pi] \to [1, 2]$ which converge L^p to $f_\infty = 1$ but do not converge pointwise almost everywhere without taking a subsequence. Let

$$\{s_j : j \in \mathbb{N}\} = \left\{\tfrac{0}{1}, \tfrac{1}{1}, \tfrac{0}{2}, \tfrac{1}{2}, \tfrac{2}{2}, \tfrac{0}{4}, \tfrac{1}{4}, \tfrac{2}{4}, \tfrac{3}{4}, \ldots\right\}$$

and

$$\{\delta_j : j \in \mathbb{N}\} = \left\{\tfrac{1}{1}, \tfrac{1}{1}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \ldots\right\}.$$

Let

$$f_j(r) = \begin{cases} h((r - s_j)/\delta_j), & r \in [s_j - \delta_j, s_j + \delta_j], \\ 1, & \text{elsewhere,} \end{cases}$$

where h is a smooth even function such that $h(-1) = 1$ with $h'(-1) = 0$, increasing up to $h(0) = h_0 \in (1, 2]$ and then decreasing back down to $h(1) = 1$, $h'(1) = 0$. Note that this defines a sequence of smooth Riemannian metrics, g_j, as in (2), with distances, d_j, as in (4) on the manifolds,

$$M_j = [-\pi, \pi] \times_{f_j} \Sigma \quad \text{or} \quad N_j = \mathbb{S}^1 \times_{f_j} \Sigma, \tag{12}$$

for any fixed Riemannian manifold Σ. Consider also M_∞ and N_∞ defined as above with $f_\infty(r) = 1$ for all r. Here we have

$$f_j \to f_\infty = 1 \quad \text{in } L^p \text{ but not pointwise} \tag{13}$$

and yet $M_j \to M_\infty$ and $N_j \to N_\infty$ in both the GH and $\mathcal{F}$ sense.

Proof. First we check that f_j converges in L^p but not pointwise almost everywhere. To this end we check that

$$\left(\int_{-\pi}^{\pi} |f_j - 1|^p \, dr\right)^{1/p} = \left(\int_{s_j - \delta_j}^{s_j + \delta_j} |h_j - 1|^p \, dr\right)^{1/p} = (2\delta_j)^{1/p} \to 0$$

since $|h_j - 1|^p \leq 1$ by construction. Of course we do not find pointwise convergence for any $r \in [-\pi, \pi]$ since for every choice of $J > 0$ one can find a $j_1 \geq J$ so that $f_{j_1}(r) = 0$ and another $j_2 \geq J$ so that $f_{j_2}(r) > 0$.

The proof of the Gromov–Hausdorff and intrinsic flat convergence follows almost exactly as in Example 3.7 except that we must choose J_δ and T_δ differently. We skip this proof since the convergence follows from Theorem 1.1 anyway. □

3G. ***Many ridges.*** Here we see a sequence of spaces which have f_j converging to $f_\infty = 1$ in the L^p sense and $f_j \geq 1$. The sequence converges pointwise to a nowhere continuous function. Nevertheless by Theorem 1.1 there is a GH and a SWIF limit without taking a subsequence and indeed the limit is the isometric product space.

Example 3.9. We first construct a classical sequence of smooth functions $f_j : [-\pi, \pi] \to [1, 2]$ as in Figure 3 which converge L^p to $f_\infty = 1$ but do not converge pointwise almost everywhere without taking a subsequence. Let

$$\begin{aligned} S &= \left\{s_{i,j} = -\pi + 2\pi i/2^j : i = 1, 2, \ldots, (2^j - 1),\ j \in \mathbb{N}\right\} \\ &= \left\{-\pi + \tfrac{2\pi}{2}, -\pi + \tfrac{2\pi}{4}, -\pi + \tfrac{2\pi 2}{4}, -\pi + \tfrac{2\pi 3}{4}, -\pi + \tfrac{2\pi}{8}, \ldots\right\} \end{aligned}$$

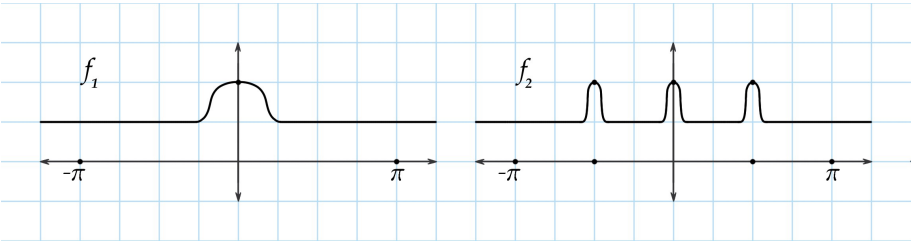


Figure 3. The warping functions of Example 3.9.

which is dense in $[-\pi, \pi]$ and

$$\{\delta_j = (\tfrac{1}{2})^{2j} : j \in \mathbb{N}\} = \{\tfrac{1}{4}, \tfrac{1}{16}, \tfrac{1}{32}, \ldots\}.$$

Let

$$f_j(r) = \begin{cases} h((r - s_{i,j})/\delta_j), & r \in [s_{i,j} - \delta_j, s_{i,j} + \delta_j] \text{ for } i = 1, \ldots, 2^j - 1, \\ 1, & \text{elsewhere,} \end{cases}$$

where h is a smooth even function such that $h(-1) = 1$ with $h'(-1) = 0$, increasing up to $h(0) = h_0 \in (1, 2]$ and then decreasing back down to $h(1) = 1$ with $h'(1) = 0$. Note that this defines a sequence of smooth Riemannian metrics, g_j, as in (2), with distances, d_j, as in (4) on the manifolds,

$$M_j = [-\pi, \pi] \times_{f_j} \Sigma \quad \text{or} \quad N_j = \mathbb{S}^1 \times_{f_j} \Sigma$$

for any fixed Riemannian manifold Σ. Consider also M_∞ and N_∞ defined as above with $f_\infty(r) = 1$ for all r. Here we have

$$f_j \to f_\infty = 1 \quad \text{in } L^p \text{ but not pointwise}$$

and yet $M_j \to M_\infty$ and $N_j \to N_\infty$ in both the GH and $\mathcal{F}$ sense.

Proof. First we check that f_j converges in L^p

$$\left(\int_{-\pi}^{\pi} |f_j - 1|^p \, dr\right)^{1/p} = \left(\sum_{i=1}^{2^j - 1} \int_{s_{i,j} - \delta_j}^{s_{i,j} + \delta_j} |f_j - 1|^p \, dr\right)^{1/p} = ((2^j - 1)(2\delta_j))^{1/p}$$
$$= \left((2^j - 1)(\tfrac{1}{2})^{2j}\right)^{1/p} \to 0.$$

Next observe that f_j converges pointwise on S to h_0 and pointwise to 1 elsewhere. Since S is dense and $h_0 > 1$ the pointwise limit is continuous nowhere.

The proof of the Gromov–Hausdorff and intrinsic flat convergence follows almost exactly as in Example 3.7 except that we must choose J_δ and T_δ differently. We skip this proof since the convergence follows from Theorem 1.1 anyway. □

3H. ***Converging to Euclidean-taxi spaces.*** In Theorem 1.1 we will prove that if $f_j \geq 1$ and $f_j \to 1$ in the L^p sense then we have Gromov–Hausdorff and intrinsic flat convergence to the isometric product space just as in Examples 3.7, 3.8 and 3.9. We now investigate what might happen if f_j does not converge to 1 in the L^p sense but does have a dense collection of points where f_j converges pointwise to 1. In the example below we see that this does not suffice to prove GH or intrinsic flat convergence to the isometric product space.

Here we will construct a sequence of warped product spaces with increasingly many cinches. The limit metric we obtain in this example is not a Riemannian metric but a metric of the following form:

Definition 3.10. Let M and N be product manifolds as in (1). For any $R > 1$, we define the minimized R-stretched Euclidean taxi metric (R-ET metric) between $x_1 = (s_1, \theta_1)$ and $x_2 = (s_2, \theta_2)$ to be

$$d^M_{R\text{-ET}}(x_1, x_2) = \min_{\Theta \in [0, d_\Sigma(\theta_1,\theta_2)]} \sqrt{|s_1 - s_2|^2 + R^2\Theta^2} + d_\Sigma(\theta_1, \theta_2) - \Theta,$$

$$d^N_{R\text{-ET}}(x_1, x_2) = \min_{\Theta \in [0, d_\Sigma(\theta_1,\theta_2)]} \sqrt{d_{\mathbb{S}^1}(s_1, s_2)^2 + R^2\Theta^2} + d_\Sigma(\theta_1, \theta_2) - \Theta.$$

Note that the R-ET metric is smaller than the isometric product metric with the θ direction scaled by R (achieved at $\Theta = d_\Sigma(\theta_1, \theta_2)$), and it is also smaller than the taxi product (achieved at $\Theta = 0$). One may view the R-ET metric as an infimum over lengths of all curves which are partly line segments of the form $\theta = ms + \theta_0$ (whose lengths are measured by stretching the Euclidean metric by R in the θ direction) and partly vertical segments purely in the θ direction (whose lengths are not rescaled). Without stretching, taking $R = 1$, we see the minimum is achieved going purely diagonal with the standard Euclidean metric.

It is not immediately obvious that R-ET metrics are true metrics satisfying positivity, symmetry and the triangle inequality. We prove this in the following lemma.

Lemma 3.11. *When*

$$d_\Sigma(\theta_1, \theta_2) \leq \frac{|s_1 - s_2|}{R\sqrt{R^2 - 1}}, \tag{14}$$

the metric is an isometric product

$$d^M_{R\text{-ET}}((s_1, \theta_1), (s_2, \theta_2)) = \sqrt{|s_1 - s_2|^2 + R^2 d_\Sigma(\theta_1, \theta_2)^2}, \tag{15}$$

and otherwise the metric is a stretched taxi product:

$$d^M_{R\text{-ET}}((s_1, \theta_1), (s_2, \theta_2)) = |s_1 - s_2|\left(\frac{\sqrt{R^2 - 1}}{R}\right) + d_\Sigma(\theta_1, \theta_2). \tag{16}$$

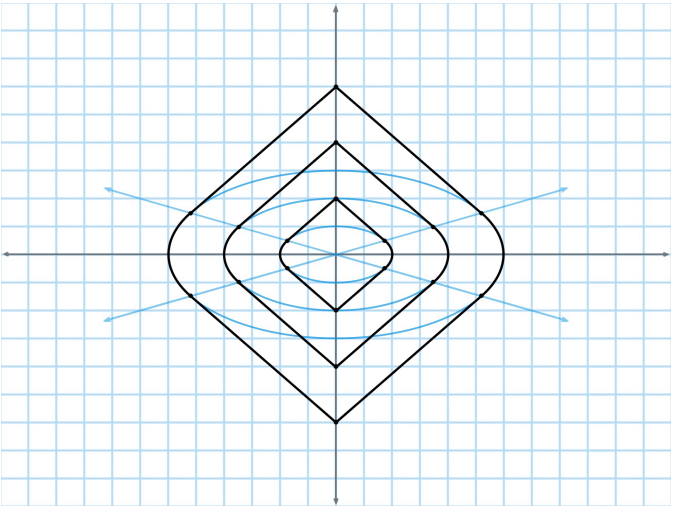


Figure 4. The concentric balls of radius $r = 2$, 4, and 6 in an R-ET space with $R = 2$ are unions of diamonds, $|s| + \frac{\sqrt{3}}{2}|\theta| < r$, and ellipses, $s^2 + 2\theta^2 < r^2$.

In fact $d^M_{R\text{-ET}}$ is a minimum of these two metrics and is a length metric whose balls are the unions of diamonds and ellipses (as in Figure 4). *It is a true metric satisfying positivity, symmetry and the triangle inequality.*

Proof. To locate the minimum in the definition of the ET metric, we take the derivative

$$\frac{d}{d\Theta}\sqrt{|s_1 - s_2|^2 + R^2\Theta^2} + d_\Sigma(\theta_1, \theta_2) - \Theta = \tfrac{1}{2}(|s_1 - s_2|^2 + R^2\Theta^2)^{-1/2}(2R^2\Theta) - 1.$$

This derivative is negative at $\Theta = 0$ so the minimum is not achieved by the taxi product metric. The derivative becomes 0 at

$$\Theta_0 = \frac{|s_1 - s_2|}{R\sqrt{R^2 - 1}} \tag{17}$$

and is then positive for $\Theta > \Theta_0$. If (14) holds then Θ_0 does not lie in $(0, d_\Sigma(\theta_1, \theta_2))$, so the minimum is achieved at $\Theta = d_\Sigma(\theta_1, \theta_2)$ and we have (15).

Otherwise, the minimum is achieved at Θ_0. Since

$$R^2\Theta_0^2 = |s_1 - s_2|^2/(R^2 - 1) \quad \text{and} \quad 1 + (1/(R^2 - 1)) = R^2/(R^2 - 1)$$

we have

$$\begin{aligned} d^M_{R\text{-ET}}((s_1,\theta_1),(s_2,\theta_2)) &\leq \sqrt{|s_1-s_2|^2+R^2\Theta_0^2}+d_\Sigma(\theta_1,\theta_2)-\Theta_0 \\ &= \frac{|s_1-s_2|\cdot|R|}{\sqrt{R^2-1}}+d_\Sigma(\theta_1,\theta_2)-\frac{|s_1-s_2|}{R\sqrt{R^2-1}} \\ &= \frac{|s_1-s_2|(R^2-1)}{R\sqrt{R^2-1}}+d_\Sigma(\theta_1,\theta_2) \\ &= |s_1-s_2|\frac{\sqrt{R^2-1}}{R}+d_\Sigma(\theta_1,\theta_2). \end{aligned}$$

Thus we have (16).

We also see that $d^M_{R\text{-ET}}((s_1,\theta_1),(s_2,\theta_2))$ is the minimum of the two metrics in (15) and (16). We know that both these metrics are length metrics. Indeed the metric in (15) is the infimum of the lengths of curves, $C(t)=(s(t),\theta(t))$ where

$$L_E(C)=\int_0^1 \sqrt{s'(t)^2+R^2 g_\Sigma(\theta'(t),\theta'(t))}\,dt$$

and the metric in (16) is the infimum of the lengths of curves $C(t)=(s(t)\theta(t))$ where

$$L_T(C)=\int_0^1 |s'(t)|\frac{\sqrt{R^2-1}}{|R|}+g_\Sigma(\theta'(t),\theta'(t))^{1/2}\,dt.$$

Thus

$$d^M_{R\text{-ET}}(x_1,x_2)=\min\{\inf_C L_E(C),\inf_C L_T(C)\}=\inf_C L_{R\text{-ET}}(C),$$

where $L_{R\text{-ET}}(C)=\min\{L_E(C),L_T(C)\}$. Thus we have positivity and symmetry (which was easy to see) and now the triangle inequality as well (which was not). □

We now present our example: a sequence of warped product spaces with increasingly many cinches which converges in the uniform, GH and $\mathcal{F}$ sense to a produce space with a minimized R-stretched Euclidean taxi metric. Here we have $R=5$, but we could easily construct similar sequences converging to any R-ET metric with $R>1$.

Example 3.12. Let

$$\begin{aligned} S &= \left\{s_{i,j}=-\pi+2\pi i/2^j : i=1,2,\ldots,(2^j-1),\ j\in\mathbb{N}\right\} \\ &= \left\{-\pi+\tfrac{2\pi}{2},-\pi+\tfrac{2\pi}{4},-\pi+\tfrac{2\pi 2}{4},-\pi+\tfrac{2\pi 3}{4},-\pi+\tfrac{2\pi}{8},\ldots\right\} \end{aligned}$$

which is dense in $[-\pi,\pi]$ and

$$\left\{\delta_j=\left(\tfrac{1}{2}\right)^{2j} : j\in\mathbb{N}\right\}=\left\{\tfrac{1}{4},\tfrac{1}{16},\tfrac{1}{32},\ldots\right\}$$

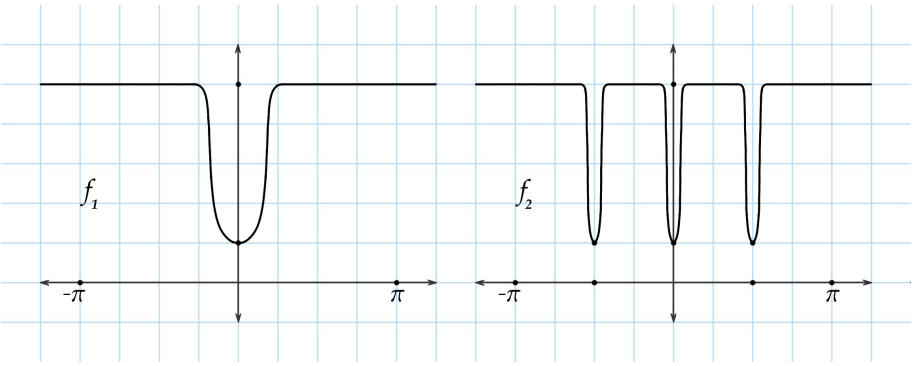


Figure 5. The warping functions of Example 3.12.

Define the functions f_j as in Figure 5 as follows:

$$f_j(r) = \begin{cases} h((r - s_{i,j})/\delta_j), & r \in [s_{i,j} - \delta_j, s_{i,j} + \delta_j] \text{ for } i = 1, \ldots, 2^j - 1, \\ 5, & \text{elsewhere,} \end{cases}$$

where h is a smooth even function such that $h(-1) = 5$ with $h'(-1) = 0$, decreasing down to $h(0) = 1$ and then increasing back up to $h(1) = 5$ with $h'(1) = 0$.

Then $f_j(r) \geq 1$ converges pointwise to 1 on the dense set, S.

If we define M_j and N_j as in (1) then they do not converge to isometric products with warping function 1. Instead they converge in the GH and $\mathcal{F}$ sense to a product manifold with an R-ET metric with $R = 5$.

Proof. First we check that $f_j \to 5$ in L^p by using the fact that $|f_j - 5|^p \leq 4^p$:

$$\begin{aligned} \left(\int_{-\pi}^{\pi} |f_j - 5|^p \, dr \right)^{1/p} &= \left(\sum_{i=1}^{2^j - 1} \int_{s_{i,j} - \delta_j}^{s_{i,j} + \delta_j} |f_j - 5|^p \, dr \right)^{1/p} \\ &\leq ((2^j - 1)(2\delta_j)4^p)^{1/p} \\ &= 4\left((2^j - 1)\left(\tfrac{1}{2}\right)^{2j}\right)^{1/p} \to 0. \end{aligned}$$

Now observe that since

$$1 \leq f_j(r) \leq 5 \quad \text{for all } r \in [-\pi, \pi] \tag{18}$$

we have

$$d_1(p, q) \leq d_j(p, q) \leq 5 d_1(p, q),$$

where d_1 is the warped product metric with warping function 1. Thus by [Huang et al. 2017], a subsequence of the warped product manifolds converges in the uniform, GH and intrinsic flat sense to some limit metric space with limit metric

d_∞

$$d_1(p,q) \leq d_\infty(p,q) \leq 5d_1(p,q) \quad \text{for all } p,q.$$

We will show that the pointwise limit of the d_j is $d_{5\text{-ET}}$, thus proving that the original sequence of warped product manifolds converges in the uniform, GH and intrinsic flat sense to the Euclidean taxi space.

Let us consider an arbitrary pair of points, $x_i = (s_i, \theta_i)$. If $\theta_1 = \theta_2$ then

$$d_j(x_1, x_2) = |s_1 - s_2| = d_{5\text{-ET}}(s_1, s_2).$$

In general, if $\theta_1 \neq \theta_2$ let $s'_{i,j} \in f_j^{-1}(1)$ with

$$|s'_{i,j} - s_i| < 2\pi/2^j, \quad x'_{i,j} = (s'_{i,j}, \theta_i).$$

By the triangle inequality applied two ways we have

$$\begin{aligned} |d_j(x_1,x_2) - d_j(x'_{1,j}, x'_{2,j})| &\leq d_j(x_1, x'_{1,j}) + d_j(x'_{2,j}, x_2) \\ &\leq |s_1 - s'_{1,j}| + |s'_{2,j} - s_2| < 4\pi/2^j \end{aligned} \tag{19}$$

and

$$\begin{aligned} |d_{5\text{-ET}}(x_1,x_2) - d_{5\text{-ET}}(x'_{1,j}, x'_{2,j})| &\leq d_{5\text{-ET}}(x_1, x'_{1,j}) + d_{5\text{-ET}}(x'_{2,j}, x_2) \\ &\leq |s_1 - s'_{1,j}| + |s'_{2,j} - s_2| < 4\pi/2^j. \end{aligned} \tag{20}$$

Recall that to complete the proof we must prove the pointwise limit:

$$\lim_{j\to\infty} d_j(x_1,x_2) = d_{5\text{-ET}}(x_1,x_2).$$

By (19) we need only show

$$\lim_{j\to\infty} d_j(x'_{1,j}, x'_{2,j}) = d_{5\text{-ET}}(x_1,x_2).$$

Applying the triangle inequality again, with $x_{1,\theta,j} = (s'_{1,j}, \theta)$, where $\theta \in \Sigma$ so that $d_\Sigma(\theta_2,\theta) \in [0, d_\Sigma(\theta_1,\theta_2)]$, we have

$$\begin{aligned} d_j(x'_{1,j}, x'_{2,j}) &\leq d_j(x'_{1,j}, x_{1,\theta,j}) + d_j(x_{1,\theta,j}, x'_{2,j}) \\ &\leq d_\Sigma(\theta_1,\theta) + \sqrt{|s'_{1,j} - s'_{2,j}|^2 + 25 d_\Sigma(\theta_2,\theta)^2}, \end{aligned}$$

where we have used (18) in the last line. Since this is true for any $\theta \in \Sigma$ such that $d_\Sigma(\theta_2,\theta) \in [0, d_\Sigma(\theta_1,\theta_2)]$ we find

$$d_j(x'_{1,j}, x'_{2,j}) \leq d_{5\text{-ET}}(x'_{1,j}, x'_{2,j}).$$

Thus taking the limsup and applying (20) we have

$$\limsup_{j\to\infty} d_j(x'_{1,j}, x'_{2,j}) \leq \limsup_{j\to\infty} d_{5\text{-ET}}(x'_{1,j}, x'_{2,j}) = d_{5\text{-ET}}(x_1,x_2). \tag{21}$$

So now we need only show

$$\liminf_{j\to\infty} d_j(x'_{1,j}, x'_{2,j}) \geq d_{5\text{-ET}}(x_1, x_2). \tag{22}$$

By (20) we need only show

$$\liminf_{j\to\infty} \big(d_j(x'_{1,j}, x'_{2,j}) - d_{5\text{-ET}}(x'_{1,j}, x'_{2,j})\big) \geq 0. \tag{23}$$

If $s'_{1,j} = s'_{2,j}$ then

$$d_j(x'_{1,j}, x'_{2,j}) \geq d_\Sigma(\theta_1, \theta_2) = d_{5\text{-ET}}(x'_{1,j}, x'_{2,j}).$$

If $s'_{1,j} \neq s'_{2,j}$, then the L_j shortest path, $C_j(t) = (r(t), \theta(t))$, from $x'_{1,j}$ to $x'_{2,j}$ must pass from one valley over to the other, possibly passing through many valleys in between. Observe that

$$d_j(x'_{1,j}, x'_{2,j}) = L_j(C_j) = L_j(C_j \cap f^{-1}(5)) + L_j(C_j \setminus f^{-1}(5)). \tag{24}$$

The segments of C_j which intersect $f_j^{-1}(5)$ lie in an product space warped by the constant function 5 so

$$L_j(C_j \cap f^{-1}(5)) = \sqrt{R_j^2 + 25\Theta_j^2}, \tag{25}$$

where R_j is the sum of changes in r on these segments and where Θ_j is the sum of distances in Σ between the theta values of the endpoints of these segments.

Let $R_0 = |s_1 - s_2|$ which is the total change in r along C_j. By the definition of δ_j,

$$2^j \delta_j = 2^j \left(\tfrac{1}{2}\right)^{2j} \to 0.$$

Since we have at most 2^j intervals where $f_j < 5$, we see that as

$$\lim_{j\to\infty} R_0 - R_j = 0, \tag{26}$$

the total change in r for the segments in $C_j \setminus f^{-1}(5)$ is converging to 0.

Let $\Theta_0 = d_\Sigma(\theta_1, \theta_2)$. Then $\Theta_0 - \Theta_j$ is the sum of distances in Σ between the theta values of the endpoints of the segments in $C_j \setminus f^{-1}(5)$. Since the warping factors $f_j(r) \geq 1$ everywhere, the distance between the endpoints of each segment is $\geq$ distance in Σ between the theta values of the endpoints of the segment. Thus

$$L_j(C_j \setminus f^{-1}(5)) \geq \Theta_0 - \Theta_j. \tag{27}$$

Combining this together with (24) and (25) we have

$$\begin{aligned} d_j(x'_{1,j}, x'_{2,j}) = L_j(C_j) &\geq \sqrt{R_j^2 + 25\Theta_j^2} + \Theta_0 - \Theta_j \\ &\geq \inf_{\Theta\in[0, d_\Sigma(\theta_1,\theta_2)]} \sqrt{R_j^2 + 25\Theta^2} + \Theta_0 - \Theta. \end{aligned} \tag{28}$$

Since

$$\lim_{j\to\infty}\left(\inf_{\Theta\in[0,d_\Sigma(\theta_1,\theta_2)]}\sqrt{R_j^2+25\Theta^2}+\Theta_0-\Theta\right)=\lim_{j\to\infty}d_{5\text{-ET}}(x'_{1,j},x'_{2,j}) \tag{29}$$

we are done by combining (28) and (29) which shows (23). □

Remark 3.13. If we take the isometric product of Example 3.12 with a standard circle, $\bar{N}_j^3=N_j^2\times\mathbb{S}^1$, $\Sigma=\mathbb{S}^1$, then we have a sequence of 3-manifolds satisfying all the hypotheses of the scalar compactness conjecture of Gromov and Sormani (see [Gromov 2018]), recently proved in the rotationally symmetric case by Park, Tian, and Wang [Park et al. 2018],

$$\mathrm{Vol}(\bar{N}_j)\le 5\,\mathrm{Vol}(\mathbb{T}^3),\quad \mathrm{Diam}(\bar{N}_j)\le 5\,\mathrm{Diam}(\mathbb{T}^3),\quad \mathrm{minA}(\bar{N}_j)\ge\mathrm{minA}(\mathbb{T}^3),$$

except for the scalar curvature bound. Therefore, this example demonstrates that the conclusion of the scalar compactness conjecture, that the SWIF limit have Euclidean tangent cones almost everywhere, requires the scalar curvature bound. We note that the volume and diameter bound follow since $f_j\le 5$ and the minA bound follows since $f_j\ge 1$.

4. Proof of the main theorem

The goal of this section is to prove our main theorem, Theorem 1.1.

In this theorem, $M_j=[r_0,r_1]\times_{f_j}\Sigma$, where Σ is an $n-1$ dimensional manifold including also M_j without boundary that have f_j periodic with period r_1-r_0 as in (1). We assume that the warping factors, $f_j\in C^0([r_0,r_1])$, satisfy the following:

$$0<f_\infty-\frac{1}{j}\le f_j(r)\le K\quad\text{and}\quad f_j(r)\to f_\infty(r)\quad\text{in } L^2,$$

where $f_\infty\in C^0([r_0,r_1])$.

The proof of Theorem 1.1 proceeds as follows. In Lemma 4.1 we use the C^0 lower bound to show that

$$\liminf_{j\to\infty}d_j(p,q)\ge d_\infty(p,q)\quad\text{pointwise.}$$

We use the L^2 convergence of $f_j\to f_\infty$ in Lemmas 4.3 and 4.6, combined with the estimate of Lemma 4.4, to show that the lengths of fixed curves with respect to M_j and M_∞ converge. We apply this result to a fixed geodesic with respect to g_∞, to prove that

$$\limsup_{j\to\infty}d_j(p,q)\le d_\infty(p,q)\quad\text{pointwise.}$$

Thus in Proposition 4.8 we have the pointwise limit

$$\lim_{j\to\infty} d_j(p,q) = d_\infty(p,q).$$

To complete the proof of uniform, GH and SWIF convergence using Theorem 2.4, as is done in the examples in Section 3, we need uniform bounds on d_j proven in Lemma 2.3.

4A. ***Assuming a C^0 lower bound.*** We have seen in Section 3 that in order to get Gromov–Hausdorff convergence to agree with L^2 convergence we will need a C^0 lower bound on f_j and so now we see the consequence of this assumption for the distance between points.

Lemma 4.1. *Let $p, q \in [r_0, r_1] \times \Sigma$ and assume that*

$$f_j(r) \geq f_\infty - \frac{1}{j} > 0, \quad \operatorname{Diam}(M_j) \leq D.$$

Then

$$\liminf_{j\to\infty} d_j(p,q) \geq d_\infty(p,q)$$

and furthermore we find the uniform estimate

$$d_{g_j}(p,q) - d_{g_\infty}(p,q) \geq -\frac{\sqrt{2}\max_{[r_0,r_1]}\sqrt{f_\infty}D}{\min_{[r_0,r_1]} f_j(r)\sqrt{j}}.$$

Proof. Let $C_j(t) = (r_j(t), \theta_j(t))$ be the absolutely continuous curve in M_j, parametrized so that $|C_j|_{g_j} = 1$ a.e., realizing the distance between p and q. Then compute

$$\begin{aligned}
d_{g_j}(p,q) &= \int_0^{L_j(C_j)} \sqrt{r_j(t)^2 + f_j(r_j(t))^2|\theta_j'(t)|^2}\,dt \\
&\geq \int_0^{L_j(C_j)} \sqrt{r_j(t)^2 + \left(f_\infty(r_j(t)) - \tfrac{1}{j}\right)^2|\theta_j'(t)|^2}\,dt \\
&= \int_0^{L_j(C_j)} \Big(r_j(t)^2 + f_\infty(r_j(t))^2|\theta_j'(t)|^2 \\
&\qquad - \left((2/j)\,f_\infty(r_j(t))|\theta_j'(t)|^2 - (1/j^2)|\theta_j'(t)|^2\right)\Big)^{1/2} dt
\end{aligned} \tag{30}$$

Now we use the inequality $\sqrt{|a-b|} \geq |\sqrt{a} - \sqrt{b}| \geq \sqrt{a} - \sqrt{b}$ in succession, employing the fact that the last integrand in (30) is positive and the square roots

that follow are of positive quantities by the assumptions of the lemma.

$$
\begin{aligned}
&d_{g_j}(p,q)\\
&\geq \int_0^{L_j(C_j)} \left| \sqrt{r_j(t)^2 + f_\infty(r_j(t))^2 |\theta_j'(t)|^2} - \tfrac{1}{\sqrt{j}} |\theta_j'(t)| \sqrt{\left(2 f_\infty(r_j(t)) - \tfrac{1}{j}\right)} \right| dt\\
&\geq \int_0^{L_j(C_j)} \sqrt{r_j(t)^2 + f_\infty(r_j(t))^2 |\theta_j'(t)|^2}\, dt\\
&\qquad\qquad - \tfrac{1}{\sqrt{j}} \int_0^{L_j(C_j)} |\theta_j'(t)| \sqrt{\left(2 f_\infty(r_j(t)) - \tfrac{1}{j}\right)}\, dt\\
&\geq L_{g_\infty}(C_j) - \tfrac{1}{\sqrt{j}} \int_0^{L_j(C_j)} |\theta_j'(t)| \sqrt{\left(2 f_\infty(r_j(t)) - \tfrac{1}{j}\right)}\, dt
\end{aligned}
$$

Now we notice that

$$
\sqrt{f_j'(t)^2 + f_j(r_j(t))^2 |\theta_j'(t)|^2} = 1 \quad \text{a.e.} \quad \Rightarrow \quad |\theta_j'(t)| \leq \frac{1}{\min f_j} \quad \text{a.e.}
$$

which allows us to compute

$$
d_{g_j}(p,q) \geq d_{g_\infty}(p,q) - \frac{\sqrt{2} \max_{[r_0,r_1]} \sqrt{f_\infty} D}{\min_{[r_0,r_1]} f_j(r) \sqrt{j}},
$$

where the diameter bound from the hypotheses is used to conclude that $L_j(C_j) \leq D$. The desired result follows by taking limits. □

4B. ***L^2 convergence and convergence of lengths.*** In this section we would like to observe the consequence of L^2 convergence of $f_j \to f_\infty$ for convergence of lengths of curves and distances between points in M_j culminating in an estimate on the pointwise limsup of the distance functions (Proposition 4.7).

We start by proving we have uniform bounds on the diameter.

Lemma 4.2. *If $\|f_j - f_\infty\|_{L^2} \leq \delta_j$ and M_j are warped products as in* (1) *then*

$$
\operatorname{Diam}(M_j) \leq 2|r_1 - r_0| + \left(\|f_\infty\|_{C_0} + \delta_j / \sqrt{r_1 - r_0} \right) \operatorname{Diam}(\Sigma) \tag{31}
$$

Proof. Let $p, q \in M_j$. Recall that the distance between these points is the infimum over lengths of all curves. For any $r \in [r_0, r_1]$ we can take a first path from p radially to the level r, then a second path around that level r, and then a third path from that level to q. The first and third paths each have length $\leq |r_1 - r_0|$, and the middle path has length bounded above by the diameter of the level. Thus we have

$$
\begin{aligned}
d_j(p,q) &\leq 2|r_1 - r_0| + f_j(r) \operatorname{Diam}(\Sigma)\\
&\leq 2|r_1 - r_0| + \left(f_\infty(r) + |f_j(r) - f_\infty(r)| \right) \operatorname{Diam}(\Sigma).
\end{aligned}
$$

Choosing an r such that

$$|f_j(r) - f_\infty(r)|^2 \le \frac{1}{r_1 - r_0} \int |f_j(s) - f_\infty(s)|^2 \, ds$$

we have

$$|f_j(r) - f_\infty(r)| \le \frac{\|f_j - f_\infty\|_{L_2}}{\sqrt{r_1 - r_0}}$$

and $f_\infty(r) \le \|f_\infty\|_{C_0}$. □

Recall that in warped product manifolds with continuous warping functions we have absolutely continuous curves whose length achieves the distance between two points (Remark 2.2).

We next consider the length of a fixed curve which is monotone in r.

Lemma 4.3. *Fix an absolutely continuous curve $C(t) = (r(t), \theta(t))$, $t \in [0, 1]$, which is monotone in r. If $\|f_j - f_\infty\|_{L^2} \le \delta = \delta_j$ and M_j are warped products as in* (1) *then*

$$|L_j(C) - L_\infty(C)| \le (\delta^2 + 4\|f_\infty\|_{L^2}^2)\delta^{1/2}\Theta(C)$$

where

$$\Theta(C) = \left(\int_{r(0)}^{r(1)} |\theta'(r)|^2 \, dr \right)^{1/2}. \tag{32}$$

Note also that

$$\|f_j + f_\infty\|_{L^2}^2 \le (\delta + 2\|f_\infty\|_{L^2})^2.$$

If C is not monotone in r but one knows it has at most N monotone subsegments then we can sum up the segments applying this lemma to each subsegment.

Proof. Since $C(t) = (r(t), \theta(t))$ is such that $r'(t) > 0$ everywhere then we can reparametrize so that $r(t) = r$. Now by comparing two lengths and taking advantage of the inequality $\sqrt{|a - b|} \ge |\sqrt{a} - \sqrt{b}|$ we find

$$\begin{aligned}
|L_j(C) - L_\infty(C)| &\le \int_{r(0)}^{r(1)} \left| \sqrt{1 + f_j^2(r))\theta'(r)^2} - \sqrt{1 + f_\infty^2(r)\theta'(r)^2} \right| dr \\
&\le \int_{r(0)}^{r(1)} \sqrt{|f_j^2(r) - f_\infty^2(r)|}\, |\theta'(r)| \, dr \\
&\le \left(\int_{r(0)}^{r(1)} |f_j^2(r) - f_\infty^2(r)| \, dr \right)^{1/2} \left(\int_{r(0)}^{r(1)} |\theta'(r)|^2 \, dr \right)^{1/2},
\end{aligned}$$

where we used Holder's inequality in the last line.

Now we notice that

$$\begin{aligned}|f_j^2 - f_\infty^2| &= |f_j^2 - f_j f_\infty + f_j f_\infty - f_\infty^2| \\ &= |f_j(f_j - f_\infty) + f_\infty(f_j - f_\infty)| \\ &= |(f_j + f_\infty)(f_j - f_\infty)| = |f_j + f_\infty|\,|f_j - f_\infty|.\end{aligned}$$

Combining this with Hölder's Inequality we obtain

$$|L_j(C) - L_\infty(C)| \leq \left(\int_{r(0)}^{r(1)} |f_j + f_\infty|^2\, dr\right)^{1/4} \left(\int_{r(0)}^{r(1)} |f_j - f_\infty|^2\, dr\right)^{1/4} \Theta(C).$$

Lastly, we notice that

$$\begin{aligned}\|f_j + f_\infty\|_{L^2}^2 &= \|f_j - f_\infty + 2f_\infty\|_{L^2} \\ &\leq (\|f_j - f_\infty\|_{L^2} + 2\|f_\infty\|_{L^2})^2 \leq (\delta + 2\|f_\infty\|_{L^2})^2\end{aligned}$$

which gives us the desired uniform bound. □

Now that we have obtained a bound on fixed geodesics which are monotone in r we would like to gain some control on the term $\Theta(C)$ from Lemma 4.3 in the case where C is a fixed geodesic with respect to the metric g_j. We note that we will use Lemma 4.4 only in the case where C is a fixed geodesic with respect to g_∞ which is monotone in r but we state it in more generality below since it could be useful for future results.

Lemma 4.4. *Let M_j be a warped product manifold as in* (1). *Let $C_j(t) = (r(t), \theta(t))$ be a unit speed absolutely continuous geodesic in M_j which is nondecreasing in r and define*

$$m_j = \min_{r \in [r_0, r_1]} f_j(r) > 0.$$

Then Θ of (32) *satisfies*:

$$\Theta(C_j) \leq \frac{\sqrt{n-1}\, L_j(C_j)^{1/2}}{m_j}.$$

Proof. We can estimate $\Theta(C_j)$ by rewriting the line integral which defines $\Theta(C_j)$:

$$\Theta(C_j) = \left(\int_{r(0)}^{r(1)} |\vec{\theta}'(r)|^2\, dr\right)^{1/2} = \left(\int_0^{L_j(C_j)} |\vec{\theta}'(t)|^2 r'(t)\, dt\right)^{1/2}.$$

Now by the assumption that $|C_j'|_{g_j} = \sqrt{r'(t)^2 + f_j(r(t))^2 |\vec{\theta}_j'(t)|^2} = 1$ a.e. and $r'(t) > 0$ we find that $0 < r'(t) \leq 1$ which yields

$$\Theta(C_j) \leq \left(\int_0^{L_j(C_j)} |\vec{\theta}'(t)|^2\, dt\right)^{1/2}.$$

Note that $|C_j'|_{g_j} = \sqrt{r'(t)^2 + f_j(r(t))^2 |\vec{\theta}_j'(t)|^2} = 1$ a.e. implies that $|\vec{\theta}_j'(t)| \leq 1/f_j$ a.e. which yields the estimate

$$\Theta(C_j) \leq \left(\int_0^{L_j(C_j)} \frac{1}{f_j(r(t))^2}\, dt \right)^{1/2} \leq \frac{L_j(C_j)^{1/2}}{m_j}. \qquad \square$$

Corollary 4.5. *If the length minimizing absolutely continuous geodesic between $p, q \in M$ with respect to g_∞ is monotone in r and we let $\delta = \|f_j - f_\infty\|_{L^2}$ and $m_\infty = \min_{r \in [r_0, r_1]} f_\infty(r) > 0$ then we find the uniform estimate*

$$d_{g_j}(p,q) - d_{g_\infty}(p,q) \leq (\delta^2 + 4\|f_\infty\|_{L^2}^2)\delta^{1/2} \frac{\sqrt{n}\,\mathrm{Diam}(M_\infty)}{m_\infty}.$$

Proof. We note that by the fact that C is the length minimizing geodesic between $p, q \in M$ with respect to g_∞ we find

$$d_{g_j}(p,q) - d_{g_\infty}(p,q) \leq L_j(C) - L_\infty(C).$$

Now if we combine Lemmas 4.2, 4.3 and 4.4 then we find

$$d_{g_j}(p,q) - d_{g_\infty}(p,q) \leq (\delta^2 + 4\|f_\infty\|_{L^2}^2)\delta^{1/2} \frac{\sqrt{n}\,\mathrm{Diam}(M_\infty)}{m_\infty},$$

where $\delta = \|f_j - f_\infty\|_{L^2}$ and $m_\infty = \min_{r \in [r_0, r_1]} f_\infty(r) > 0$. $\square$

The uniform control of Corollary 4.5 will be used in the proof of Theorem 1.1 below. Now we would like to control the length of geodesics with respect to g_∞ which are constant in r.

Lemma 4.6. *Let $p, q \in [r_0, r_1] \times \Sigma$ and assume that the absolutely continuous geodesic C between p and q with respect to g_∞ is parametrized as $C = (\hat{r}, \theta(t))$, $t \in [0,1]$, for some fixed $\hat{r} \in [r_0, r_1]$. If $f_j \to f_\infty$ in L^2 then*

$$\limsup_{j \to \infty} d_{g_j}(p,q) \leq d_{g_\infty}(p,q).$$

Moreover, we can find an approximating curve C_j^ϵ between p and q so that

$$L_j(C_j^\epsilon) \leq 4\delta_j^\epsilon + L_\infty(C) + \epsilon d_\sigma(\theta(0), \theta(1)),$$

where

$$\delta_j^\epsilon \leq \frac{|f_j - f_\infty|_{L^2}^2}{\epsilon^2}.$$

Proof. Since $f_j \to f_\infty$ in L^2, if we define

$$S_\epsilon^j = \{x \in [r_0, r_1] : |f_j(x) - f_\infty(x)| \geq \epsilon\}$$

then we know that there exists a $\delta_j > 0$ such that $|S^j_\epsilon| \leq \delta_j$, where $\delta_j \to 0$ as $j \to \infty$. This follows since if $|S^j_\epsilon| \geq c > 0$ then

$$\int_{-\pi}^{\pi} |f_j - f_\infty|^2 \, dr \geq \int_{S^j_\epsilon} |f_j - f_\infty|^2 \, dr \geq c\epsilon^2$$

which leads to a contradiction. In fact,

$$\begin{aligned}\epsilon |S^\epsilon_j| \leq \int_{S^\epsilon_j} |f_j - f_\infty| \, dr &\leq |S^\epsilon_j|^{1/2} \left(\int_{S^\epsilon_j} |f_j - f_\infty|^2 \, dr \right)^{1/2} \\ &\leq |S^\epsilon_j|^{1/2} \left(\int_{-\pi}^{\pi} |f_j - f_\infty|^2 \, dr \right)^{1/2},\end{aligned}$$

which implies

$$\delta_j \leq \frac{|f_j - f_\infty|^2_{L^2}}{\epsilon^2}.$$

This implies that we can choose an $r_j \in (\hat{r}, \hat{r} + 2\delta_j)$ or $r_j \in (\hat{r} - 2\delta_j, \hat{r})$ so that $|f_j(r_j) - f_\infty(r_j)| \leq \epsilon$ and so by combining with Lemmas 3.2 and 3.6 we find a curve C^ϵ_j between p and q such that

$$\begin{aligned}d_{g_j}(p, q) &\leq L_j(C^\epsilon_j) \\ &\leq 4\delta_j + f_j(r_j) \, d_\sigma(\theta(0), \theta(1)) \\ &\leq 4\delta_j + f_\infty(r_j) \, d_\sigma(\theta(0), \theta(1)) + |f_j(r_j) - f_\infty(r_j)| d_\sigma(\theta(0), \theta(1)).\end{aligned}$$

Now by taking limits as $j \to \infty$ and using that f_∞ is continuous we find

$$\limsup_{j \to \infty} d_{g_j}(p, q) \leq f_\infty(\hat{r}) \, d_\sigma(\theta(0), \theta(1)) + \epsilon d_\sigma(\theta(0), \theta(1)).$$

Since this is true for all $\epsilon > 0$ and $d_{g_\infty}(p, q) = f_\infty(\hat{r}) \, d_\sigma(\theta(0), \theta(1))$ the desired result follows. □

We now combine these lemmas into:

Proposition 4.7. *If f_j and f_∞ are positive continuous functions, $f_j \to f_\infty$ in L^2, and $M_j = M$ are warped products as in* (1) *then*

$$\limsup_{j \to \infty} d_j(p, q) \leq d_\infty(p, q) \quad \textit{pointwise.}$$

Proof. Fix p and q in $M_j = M$. Let $C(t)$ be a minimizing curve between p and q with respect to g_∞:

$$L_\infty(C) = d_\infty(p, q).$$

By Remark 2.2, C is an absolutely continuous curve. It can be broken down into possibly infinitely many segments, each of which is either monotone in r or has constant r component. Let $\mathcal{C} = \{C^\alpha : \alpha \in I\}$, where I is an indexing set, be the

segments which are constant in r with endpoints $(r^\alpha, \theta_1^\alpha), (r^\alpha, \theta_2^\alpha) \in [r_0, r_1] \times \Sigma$ then we can estimate

$$\begin{aligned} L_\infty(C) &\geq \sum_{\alpha \in I} L_\infty(C^\alpha) \\ &= \sum_{\alpha \in I} f_\infty(r^\alpha)\, d_\sigma(\theta_1^\alpha, \theta_2^\alpha) \geq \left(\min_{r \in [r_0, r_1]} f_\infty(r)\right) \textstyle\sum_{\alpha \in I} d_\sigma(\theta_1^\alpha, \theta_2^\alpha), \end{aligned}$$

and hence

$$\sum_{\alpha \in I} d_\sigma(\theta_1^\alpha, \theta_2^\alpha) \leq \frac{\mathrm{Diam}(M_\infty)}{\left(\min_{r \in [r_0, r_1]} f_\infty(r)\right)} < \infty. \tag{33}$$

Similarly, if we let $\tilde{\mathcal{C}} = \{\tilde{C}^\alpha : \alpha \in I\}$ be the collection of segments of C which are monotone in r, with endpoints $(r_1^\alpha, \theta_1^\alpha), (r_2^\alpha, \theta_2^\alpha) \in [r_0, r_1] \times \Sigma$, then

$$\begin{aligned} L_\infty(C) &\geq \sum_{\alpha \in I} L_\infty(\tilde{C}^\alpha) = \sum_{\alpha \in I} \int_{r_1^\alpha}^{r_2^\alpha} \sqrt{1 + f_\infty(r)^2 \theta'(r)^2}\, dr \\ &\geq \sum_{\alpha \in I} \int_{r_1^\alpha}^{r_2^\alpha} dr = \sum_{\alpha \in I} |r_1^\alpha - r_2^\alpha|, \end{aligned}$$

which implies

$$\sum_{\alpha \in I} |r_1^\alpha - r_2^\alpha| \leq \mathrm{Diam}(M_\infty). \tag{34}$$

So, by combining (33), (34), and Lemma 3.2 we find for any $\eta > 0$, we can choose $I_\eta \subset I,\ I \setminus I_\eta = K \in \mathbb{N}$, so that

$$\begin{aligned} &\sum_{\alpha \in I_\eta} L_\infty(\tilde{C}^\alpha) + \sum_{\alpha \in I_\eta} L_\infty(C^\alpha) \\ &\qquad \leq \sum_{\alpha \in I_\eta} |r_1^\alpha - r_2^\alpha| + 2\left(\max_{r \in [r_0, r_1]} f_\infty(r)\right) \sum_{\alpha \in I_\eta} d_\sigma(\theta_1^\alpha, \theta_2^\alpha) \leq \eta \end{aligned} \tag{35}$$

and hence by replacing all but finitely many subsegments of C with finitely many taxi minimizing curves whose g_∞ length is smaller than η we can obtain another curve $\bar{C}^\eta$ such that

$$L_\infty(\bar{C}^\eta) \leq L_\infty(C) - 2\eta.$$

This can be done so that $\bar{C}^\eta$ can be broken down into finitely many segments, each of which is either monotone in r or has constant r component. By Lemma 4.6, for each monotone segment $\bar{C}^k,\ k \in \mathbb{N},\ k \leq K$ we can find an approximating curve, $\bar{C}_j^{k,\epsilon}$, such that

$$L_j(\bar{C}_j^{k,\epsilon}) \leq 4\delta_j^\epsilon + L_\infty(\bar{C}^k) + \epsilon d_\sigma(\theta_1^k, \theta_2^k), \tag{36}$$

where $\delta_j^\epsilon \leq |f_j - f_\infty|^2_{L^2}/\epsilon^2$.

Then by Lemmas 4.3, 4.4 and 4.6 we can find a curve $\bar{C}_j^{\eta,\epsilon}$, $\epsilon > 0$ between p and q, by possibly adjusting the monotone segments as in (36), such that

$$\limsup_{j\to\infty} L_j(\bar{C}_j^{\eta,\epsilon}) \leq L_\infty(C) - 2\eta + \epsilon \frac{\mathrm{Diam}(M_\infty)}{\left(\min_{r\in[r_0,r_1]} f_\infty(r)\right)}. \tag{37}$$

Since (37) is true for all η, $d_j(p,q) \leq L_j(\bar{C}_j^{\eta,\epsilon})$ and $L_\infty(C) = d_\infty(p,q)$ we have

$$\limsup_{j\to\infty} d_j(p,q) \leq d_\infty(p,q) + \epsilon \frac{\mathrm{Diam}(M_\infty)}{\left(\min_{r\in[r_0,r_1]} f_\infty(r)\right)},$$

which is true for all $\epsilon > 0$ and hence the desired result follows. □

4C. ***Proof of Theorem 1.1.*** Recall that in the statement of Theorem 1.1 we have a sequence of warping functions $f_j(r) \geq f_\infty(r) - \frac{1}{j}$ and $f_j(r) \to f_\infty(r)$ in L^2. We will prove:

$$\lim_{j\to\infty} d_j(p,q) = d_\infty(p,q)$$

uniformly by first showing it converges pointwise on a subsequence and then applying Theorem 2.4 which implies uniform convergence, GH and $\mathcal{F}$ convergence to the same space.

Proposition 4.8. *Under the hypothesis of Theorem 1.1 we have pointwise convergence of the distance functions*:

$$\lim_{j\to\infty} d_j(p,q) = d_\infty(p,q).$$

Proof. Let $p, q \in [r_0, r_1] \times \Sigma$. Applying the C^0 lower bound and Lemma 4.1 we have

$$\liminf_{j\to\infty} d_j(p,q) \geq d_\infty(p,q).$$

Applying the L^2 upper bound and Proposition 4.7 we also have

$$\limsup_{j\to\infty} d_j(p,q) \leq d_\infty(p,q).$$

Thus we have pointwise convergence. □

Proof of Theorem 1.1. By the assumption that $0 < c \leq f_\infty - \frac{1}{j} \leq f_j \leq K$ we can use Lemma 2.3 and choose $\lambda = \max(1/\min(c,1), \max(1,K)) > 0$ so that for j large enough we find

$$\lambda \geq \frac{d_j(p,q)}{d_1(p,q)} \geq \frac{1}{\lambda},$$

where d_1 is the distance defined with warping factor 1.

Now can apply Theorem 2.4 to conclude that there exists a length metric d'_∞ and a subsequence d_{j_k} such that d_{j_k} converges uniformly to d'_∞, and hence GH and SWIF converges as well. By the pointwise convergence proven in Proposition 4.8, we know that $d'_\infty = d_\infty$ and hence d_{j_k} must uniformly converge to d_∞. Since this is true for all the subsequences, we see that d_j uniformly converges to d_∞. Appealing again to Theorem 2.4 we see it converges in the Gromov–Hausdorff and intrinsic flat sense as well. □

5. Warping functions with two variables on tori

In this section we give a short exploration of more general warped product manifolds. There are a wealth of new directions one might explore and this section demonstrates how some of our techniques do extend easily. Here we prove the following theorem:

Theorem 5.1. *Let $g_j = dx^2 + dy^2 + f_j(x, y)^2\, dz^2$ be a metric on a torus $M_j = \mathbb{S}^1 \times \mathbb{S}^1 \times_{f_j} \mathbb{S}^1$ with coordinates $(x, y, z) \in [-\pi, \pi]^3$, $f_j \in C^0([-\pi, \pi]^2)$. Assume that*

$$f_j \to f_\infty = c > 0 \quad \text{in } L^2, \qquad 0 < f_\infty - \frac{1}{j} \le f_j \le K < \infty.$$

Then M_j converges uniformly to M_∞ as well as

$$M_j \xrightarrow{\text{GH}} M_\infty, \quad M_j \xrightarrow{\mathcal{F}} M_\infty.$$

This theorem will be applied in upcoming joint work of a team of doctoral students who are working with the first author: Lisandra Hernandez-Vazquez, Davide Parise, Alec Payne, and Shengwen Wang. Various members of this team which first began working together at the Fields Institute in the Summer of 2017 will explore further theorems in this direction using similar techniques.

The proof of this theorem will be similar to the proof of Theorem 1.1, however we have some additional difficulties arising. The main difficulty is that $f_j \to f_\infty$ in $L^2([-\pi, \pi]^2)$ does not imply that $f_j \to f_\infty$ on curves and hence we will not be able to prove the corresponding results to Lemmas 4.3 and 4.4 for this setting. Instead in Lemmas 5.4, 5.5, and 5.6 we will build approximating sequences of curves to a geodesic with respect to g_∞ and show $\limsup_{j\to\infty} d_j(p, q) \le d_\infty(p, q)$. The C^0 control on f_j works similarly to Section 4 and hence we are able to show $\liminf_{j\to\infty} d_j(p, q) \ge d_\infty(p, q)$ in Lemma 5.2. This will imply pointwise convergence of distances which when combined with Theorem 2.4 will show uniform, GH and SWIF convergence, similar to the examples in Section 3.

5A. ***A lower C^0 bound.*** We now prove a lemma which shows the consequence of a C^0 lower bound which we have seen is important by the examples in Section 3.

Lemma 5.2. *Let $p, q \in M_j$ and assume that*

$$f_j(x, y) \geq f_\infty(x, y) - \frac{1}{j} > 0 \quad \textit{and} \quad \mathrm{Diam}(M_j) \leq D.$$

Then

$$\liminf_{j\to\infty} d_j(p, q) \geq d_\infty(p, q)$$

Proof. Let $C_j(t) = (x_j(t), y_j(t), z_j(t))$ be the minimizing absolutely continuous geodesic in M_j, parametrized so that $|C_j'(t)|_{g_j} = 1$ a.e., realizing the distance between p and q. Then compute

$$\begin{aligned} g_j(C_j'(t), C_j'(t)) &= x_j'(t)^2 + y_j'(t)^2 + f_j(x_j(t), y_j(t))^2 |z_j'(t)|^2 \\ &\geq x_j'(t)^2 + y_j'(t)^2 + \left(f_\infty(x_j(t), y_j(t)) - \tfrac{1}{j}\right)^2 |z_j'(t)|^2 \\ &= x_j'(t)^2 + y_j'(t)^2 + f_\infty(x_j(t), y_j(t))^2 |z_j'(t)|^2 \\ &\qquad - \left((2/j) f_\infty(x_j(t), y_j(t)) |z_j'(t)|^2 - (1/j^2)|z_j'(t)|^2\right). \end{aligned}$$

Note that the terms here are positive by the assumptions of the lemma, so that when we take the square root we can apply the inequality

$$\sqrt{|a - b|} \geq |\sqrt{a} - \sqrt{b}| \geq \sqrt{a} - \sqrt{b},$$

before integrating to obtain

$$\begin{aligned} d_{g_j}(p, q) &= \int_0^{L_j(C_j)} \sqrt{g_j(C_j'(t), C_j'(t))}\, dt \\ &\geq \int_0^{L_j(C_j)} \sqrt{x_j'(t)^2 + y_j'(t)^2 + f_\infty(x_j(t), y_j(t))^2 |z_j'(t)|^2}\, dt \\ &\quad - \int_0^{L_j(C_j)} \sqrt{(2/j) f_\infty(x_j(t), y_j(t)) |z_j'(t)|^2 - (1/j^2)|z_j'(t)|^2}\, dt \\ &\geq L_{g_\infty}(C_j) - \tfrac{1}{\sqrt{j}} \int_0^{L_j(C_j)} |z_j'(t)| \sqrt{\left(2 f_\infty(x_j(t), y_j(t)) - \tfrac{1}{j}\right)}\, dt \end{aligned}$$

Now we notice that

$$\begin{aligned} |C_j'(t)|_{g_j} = \sqrt{x_j'(t)^2 + y_j'(t)^2 + f_j(x_j(t), y_j(t))^2 |z_j'(t)|^2} = 1 \text{ a.e.} \\ \Rightarrow |z_j'(t)| \leq \frac{1}{f_j(x_j(t), y_j(t))} \text{ a.e.} \end{aligned}$$

and hence we can then conclude that

$$d_{g_j}(p, q) \geq d_{g_\infty}(p, q) - \frac{\sqrt{2} \max_{[-\pi,\pi]^2} \sqrt{f_\infty} D}{\min_{[-\pi,\pi]^2} f_j \sqrt{j}}.$$

The desired result follows by taking limits. □

We now prove that we have uniform bounds on the diameter which was used in Lemma 5.2:

Lemma 5.3. *If $\|f_j - f_\infty\|_{L^2} \leq \delta_j$ and M_j are warped products as in Theorem 5.1 then*

$$\operatorname{Diam}(M_j) \leq 4\sqrt{2}\pi + 2\pi\big(\|f_\infty\|_{C_0} + \delta_j/(2\pi)\big).$$

Proof. Let $p, q \in M_j$ with $p = (x_1, y_1, z_1)$ and $q = (x_2, y_2, z_2)$. Recall that the distance between these points is the infimum over lengths of all curves. For any $(x_0, y_0) \in [-\pi, \pi]^2$ we can take a first path from p to (x_0, y_0, z_1) which stays in a plane parallel to the $xy-$plane, then a second path from (x_0, y_0, z_1) to (x_0, y_0, z_2) parallel to the z axis, and then a third path from (x_0, y_0, z_2) to (x_2, y_2, z_2) which stays in a plane parallel to the $xy-$plane. The first and third paths each have length $\leq 2\sqrt{2}\pi$, and the middle path has length bounded above by 2π with respect to the flat metric. Thus we have

$$\begin{aligned} d_j(p, q) &\leq 4\sqrt{2}\pi + 2\pi f_j(x_0, y_0) \\ &\leq 4\sqrt{2}\pi + 2\pi\big(f_\infty(x_0, y_0) + |f_j(x_0, y_0) - f_\infty(x_0, y_0)|\big). \end{aligned}$$

Choosing an (x_0, y_0) such that

$$|f_j(x_0, y_0) - f_\infty(x_0, y_0)|^2 \leq \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |f_j(x, y) - f_\infty(x, y)|^2\, dx\, dy$$

we have

$$|f_j(x_0, y_0) - f_\infty(x_0, y_0)| \leq \frac{\|f_j - f_\infty\|_{L_2}}{2\pi}$$

and $f_\infty(x_0, y_0) \leq \|f_\infty\|_{C_0}$. □

5B. ***L^2 convergence and convergence of distances.*** In this section we will build sequences of curves whose length approximates the length of a fixed geodesic with respect to g_∞ whose warping function is a constant.

We start by approximating a geodesic which has constant z component which is simple since g_j agrees with g_∞ in the x and y directions.

Lemma 5.4. *Let $p, q \in [-\pi, \pi]^3$ so that $p = (x_1, y_1, z_0)$ and $q = (x_2, y_2, z_0)$. If $f_\infty = c > 0$ then we have that*

$$\limsup_{j\to\infty} d_j(p, q) \leq d_\infty(p, q).$$

Proof. Let γ be a minimal geodesic with respect to g_∞ from p to q. Since g_∞ is a Euclidean metric it is a straight line segment:

$$\gamma(t) = \big(x_1(1-t) + x_2 t,\, y_1(1-t) + y_2 t,\, z_0\big).$$

Note that we can choose coordinate so that this is the minimal geodesic with respect to g_∞. Then we can compute

$$d_j(p,q) \leq L_j(\gamma) = \int_0^1 \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}\,dt = d_\infty(p,q),$$

since g_j agrees with g_∞ in the x and y directions, by which the result follows by taking limits. □

We now construct a sequence of curves which approximates a fixed geodesic with respect to g_∞ which is constant in x and y.

Lemma 5.5. *Assume that $f_j \to f_\infty = c > 0$ in L^2 and let $p, q \in [-\pi,\pi]^3$ so that $p = (x_0, y_0, z_1)$ and $q = (x_0, y_0, z_2)$ then we have that*

$$\limsup_{j\to\infty} d_j(p,q) \leq d_\infty(p,q).$$

Proof. We claim that if

$$S_\epsilon^j = \left\{(x,y) \in [-\pi,\pi]^2 : |f_j(x,y) - f_\infty(x,y)| \geq \epsilon\right\}$$

then we must have that $|S_\epsilon^j| \leq \delta_j$, where $\delta_j \to 0$ as $j \to \infty$ ($|S|$ represents Lebesgue measure of $S \subset [-\pi,\pi]^2$ with respect to the Euclidean metric). If the claim were false then $|S_\epsilon^j| \geq C > 0$ and

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |f_j(x,y) - f_\infty(x,y)|^2\,dx\,dy \geq \int_{S_\epsilon^j} |f_j(x,y) - f_\infty(x,y)|^2\,dA \geq C\epsilon^2$$

which contradicts $f_j \to f_\infty$ in L^2.

Define the set

$$T_\epsilon^j = \left(B((x_0,y_0), 4\sqrt{\delta_j}) \setminus S_\epsilon^j\right) \cap [-\pi,\pi]^2.$$

Since eventually

$$\frac{|B((x_0,y_0), 4\sqrt{\delta_j})|}{4} = 4\pi\delta_j > |S_\epsilon^j|,$$

we see that T_ϵ^j is nonempty. Hence we can choose a $(x_\epsilon^j, y_\epsilon^j) \in T_\epsilon^j$.

A minimal geodesic γ from $p = (x_0, y_0, z_1)$ to $q = (x_0, y_0, z_2)$ with respect to g_∞ is purely vertical:

$$\gamma(t) = (x_0, y_0, z_0(1-t) + z_2 t),$$

where the addition is mod 2π. Note that $d_\infty(p,q) = c|z_2 - z_1|$. Let

$$p' = (x_\epsilon^j, y_\epsilon^j, z_1) \quad \text{and} \quad q' = (x_\epsilon^j, y_\epsilon^j, z_2).$$

So $d_\infty(p,p') < 4\sqrt{\delta_j}$ and $d_\infty(q,q') < 4\sqrt{\delta_j}$. Also

$$d_\infty(p,q) = c|z_2 - z_1| = d_\infty(p',q').$$

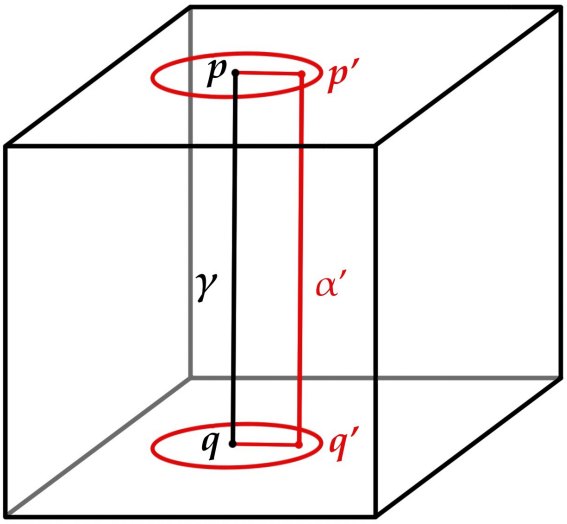


Figure 6. $\alpha' = \alpha^j_{x_j}$ approximates the curve γ between the points p and q.

We can define a curve α^j_ϵ as in Figure 6 which approximates γ. This curve runs minimally with respect to g_∞ from p to p' and then minimally to q' and then minimally to q as follows:

$$\alpha^j_\epsilon(t) = \begin{cases} (x_0(1-3t) + 3x^j_\epsilon t,\ y_0(1-3t) + 3y^j_\epsilon t,\ z_1), & 0 \le t \le \frac{1}{3}, \\ (x^j_\epsilon,\ y^j_\epsilon,\ z_1(2-3t) + z_2(3t-1)), & \frac{1}{3} \le t \le \frac{2}{3}, \\ (x^j_\epsilon(3-3t) + x_0(3t-2),\ y^j_\epsilon(3-3t)) + y_0(3t-2),\ z_2), & \frac{2}{3} \le t \le 1, \end{cases}$$

where the addition here is mod 2π.

Now we can compute

$$\begin{aligned} d_j(p,q) &\le L_j(\alpha^j_\epsilon) \\ &= \int_0^{1/3} \sqrt{|3x^j_\epsilon - 3x_0|^2 + |3y^j_\epsilon - 3y_0|^2}\, dt + \int_{1/3}^{2/3} |3z_2 - 3z_1| f_j(x^j_\epsilon, y^j_\epsilon)\, dt \\ &\qquad + \int_{2/3}^{1} \sqrt{|3x^j_\epsilon - 3x_0|^2 + |3y^j_\epsilon - 3y_0|^2}\, dt. \end{aligned}$$

Combining this with the definitions of $(x^j_\epsilon, y^j_\epsilon) \in T^j_\epsilon$ and using the continuity of f_∞ we find

$$\begin{aligned} d_j(p,q) &= 2\sqrt{|x_0 - x^j_\epsilon|^2 + |y_0 - y^j_\epsilon|^2} + f_j(x^j_\epsilon, y^j_\epsilon)|z_2 - z_1| \\ &\le 16\sqrt{\delta_j} + |f_j(x^j_\epsilon, y^j_\epsilon) - f_\infty(x^j_\epsilon, y^j_\epsilon)||z_2 - z_1| + f_\infty(x^j_\epsilon, y^j_\epsilon)|z_2 - z_1| \\ &\le 16\sqrt{\delta_j} + \epsilon|z_2 - z_1| + c|z_2 - z_1|, \end{aligned}$$

where we are using the hypothesis that $f_\infty = c > 0$.

Now by noticing that $d_\infty(p,q) = c|z_2 - z_1|$ and taking the limit as $j \to \infty$ we find

$$\limsup_{j\to\infty} d_j(p,q) \leq \epsilon|z_1 - z_0| + c|z_1 - z_0| = \epsilon|z_1 - z_0| + d_\infty(p,q)$$

and since this is true for all $\epsilon > 0$ the result follows. □

We now construct a sequence of curves which approximates a fixed geodesic with respect to g_∞, which does not fall under the hypotheses of Lemmas 5.4 or 5.5.

Lemma 5.6. *Assume that $f_j \to f_\infty = c > 0$ in L^2 and let $p, q \in [-\pi, \pi]^3$ so that $p = (x_1, y_1, z_1)$, $q = (x_2, y_2, z_2)$ and $(x_1, y_1) \neq (x_2, y_2)$ then*

$$\limsup_{j\to\infty} d_j(p,q) \leq d_\infty(p,q).$$

Proof. Without loss of generality we may assume that $y_1 \neq y_2$. Let γ be the geodesic with respect to g_∞ which runs from p to q. Since g_∞ is a Euclidean metric, we can choose coordinates on $\mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1$ such that

$$\gamma(t) = (\alpha(t), z_1(1-t) + z_2 t),$$

where the addition is mod 2π and

$$\alpha(t) = \big(x_1(1-t) + x_2 t,\, y_1(1-t) + y_2 t\big) \subset [-\pi, \pi]^2.$$

Since $g_\infty = dx^2 + dy^2 + c^2\, dz^2$, we have

$$d_\infty(p,q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

We construct a family of geodesics parallel to this geodesic running from $p' = (x_1', y_1, z_1)$ to $q' = (x_2 + x_1' - x_1, y_2, z_2)$ where $x_1' \in B(x_1, 1) \subset [-\pi, \pi]$, as follows:

$$\gamma_{x_1'}(t) = (\alpha'_{x_1}(t), z_1(1-t) + z_2 t),$$

where

$$\alpha_{x_1'}(t) = \big(x_1'(1-t) + (x_1' + x_2 - x_1)t,\, y_1(1-t) + y_2 t\big),$$

where the addition is mod 2π with values in $[-\pi, \pi)$. Observe that $\alpha : (x', t) \to (x, y)$ defined by $\alpha(x', t) = \alpha_{x'}(t)$ is

$$\alpha(x', t) = \big(x' + (x_2 - x_1)t,\, y_1 + (y_2 - y_1)t\big)$$

so

$$dx \wedge dy = (1dx' + (x_2 - x_1)\, dt) \wedge (0dx' + (y_2 - y_1)\, dt) = (y_2 - y_1)\, dx' \wedge dt. \tag{38}$$

Since $f_j \to f_\infty$ in L^2 we define

$$\bar{f}_j(x') = \int_{\alpha_{x'}} |f_j - f_\infty|^2 \, dt. \tag{39}$$

We define the set

$$S_\epsilon^j = \left\{x' \in [-\pi, \pi) : \bar{f}_j(x') \geq \epsilon\right\} \subset [-\pi, \pi),$$

and the set

$$W = \left\{\alpha_{x'}(t) : x' \in [-\pi, \pi) \text{ and } t \in [0, 1]\right\}.$$

By the definition of the line segments, $\alpha_{x_1'}$, we have $W \subset (-\pi, \pi]^2$.

Note that the set

$$T_\epsilon^j = (B(x_1, 4\delta_j) \setminus S_\epsilon^j) \subset [-\pi, \pi]$$

is nonempty where $\delta_j = |S_\epsilon^j|$. We claim $\delta_j \to 0$ as $j \to \infty$. Indeed we have

$$\epsilon |S_\epsilon^j| \leq \int_{x' \in S_\epsilon^j} \bar{f}(x') \, dx' \leq \int_{x'=-\pi}^{\pi} \bar{f}_j(x') \, dx' = \int_{x'=-\pi}^{\pi} \int_{\alpha_{x'}} |f_j - f_\infty|^2 \, dt \, dx'.$$

Applying a change of variables as in (38), we have

$$\begin{aligned} \delta_j &= (\epsilon)^{-1} \int \int_W |f_j - f_\infty|^2 |y_2 - y_1|^{-1} \, dy \, dx' \\ &\leq (\epsilon)^{-1} |y_2 - y_1|^{-1} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} |f_j - f_\infty|^2 \, dy \, dx, \end{aligned}$$

which converges to 0 by the hypothesis that $f_j \to f_\infty$ in L^2.

Since T_ϵ^j is nonempty, we can pick a $x_j \in T_\epsilon^j$. We use this point to choose

$$p' = p'_j = (x_\epsilon^j, y_\epsilon^j, z_1) \quad \text{and} \quad q' = q'_j = (x_\epsilon^j, y_\epsilon^j, z_2). \tag{40}$$

We can define a sequence of curves $\beta_{x_j}^j$ as in Figure 7 which run minimally with respect to g_∞ from p to p' and then minimally to q' and then minimally to q as follows:

$$\beta_{x_j}^j(t) = \begin{cases} (x_1(1-3t) + 3x_j t, \, y_1, \, z_1), & 0 \leq t \leq \frac{1}{3}, \\ \gamma_{x_j}(3t-1), & \frac{1}{3} \leq t \leq \frac{2}{3}, \\ ((x_j + x_2 - x_1)(3-3t) + x_2(3t-2), \, y_2, \, z_2), & \frac{2}{3} \leq t \leq 1. \end{cases}$$

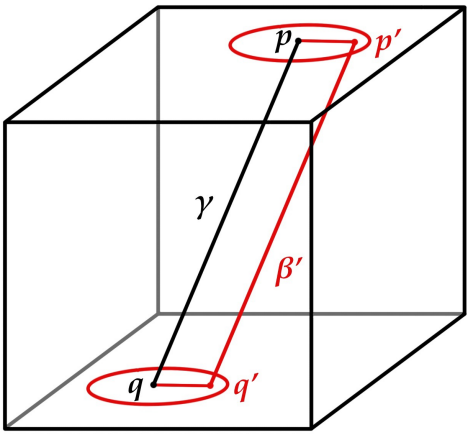


Figure 7. $\beta' = \beta^j_{x_j}$ approximates the curve γ between the points p and q.

The sequence of curves $\beta^j_{x_j}(t)$ is the approximating sequence to γ which can be used to estimate $d_j(p,q)$ as follows

$$\begin{aligned} d_j(p,q) &\leq L_j(\beta_{x_j}) \\ &= \int_0^{1/3} |3x_j - 3x_1|\, dt' + \int_{1/3}^{2/3} \sqrt{|3\Delta x|^2 + |3\Delta y|^2 + |3\Delta z|^2 f_j^2(\alpha_{x_j}(3t'-1))}\, dt' \\ &\qquad + \int_{2/3}^{1} \sqrt{|3x_2 - 3(x_j + x_2 - x_1)|^2}\, dt', \end{aligned}$$

where $\Delta x = |x_2 - x_1|$, $\Delta y = |y_2 - y_1|$, and $\Delta z = |z_2 - z_1|$. Integrating the first and last term, and taking $t = 3t' - 1$ we have

$$\begin{aligned} d_j(p,q) &\leq \left(\tfrac{1}{3} - 0\right)|3x_j - 3x_1| + \left(1 - \tfrac{2}{3}\right)\sqrt{|3x_2 - 3x_j - 3x_2 + 3x_1)|^2} \\ &\qquad + \int_0^1 \sqrt{|\Delta x|^2 + |\Delta y|^2 + |\Delta z|^2 f_j^2(\alpha_{x_j}(t))}\, dt \\ &\leq |x_j - x_1| + |x_j - x_1| + \int_0^1 \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2 f_j^2(\alpha_{x_j}(t'))}\, dt \\ &\leq 2|x_j - x_1| + \int_0^1 \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2 f_\infty^2 + \Delta z^2 (f_j^2(\alpha_{x_j}(t)) - f_\infty^2)}\, dt \\ &\leq 4\delta_j + \int_0^1 \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2 f_\infty^2}\, dt + \int_0^1 \Delta z \sqrt{f_j^2(\alpha_{x_j}(t)) - f_\infty^2}\, dt. \end{aligned}$$

Since g_∞ is Euclidean, the middle term is $d_\infty(p,q)$. Applying Hölder's inequality to the last term of yields

$$d_j(p,q) \le 4\delta_j + d_\infty(p,q) + \Delta z \left(\int_0^1 |f_j^2(\alpha_{x_j}(t)) - f_\infty^2| \, dt \right)^{1/2}. \tag{41}$$

Recall that we chose $x_j \in T_\epsilon^j$ near x so that $x_j \notin S_\epsilon^j$. Thus (39) implies that

$$\int_{\alpha_{x_j}} |f_j - f_\infty|^2 \, dt \int_0^1 |f_j(\alpha_{x_j}(t)) - f_\infty|^2 \, dt = \bar{f}_j(x_j) < \epsilon.$$

We can apply this to control the final term in (41) by factoring and the applying Hölder's inequality and the triangle inequality

$$\begin{aligned}
\left(\int_{\alpha_{x_j}} |f_j^2 - f_\infty^2| \, dt \right)^{1/2} &\le \left(\int_{\alpha_{x_j}} |f_j - f_\infty||f_j + f_\infty| \, dt \right)^{1/2} \\
&\le \left(\int_{\alpha_{x_j}} |f_j - f_\infty|^2 \, dt \right)^{1/4} \left(\int_{\alpha_{x_j}} |f_j + f_\infty|^2 \, dt \right)^{1/4} \\
&\le \epsilon^{1/4} \left(\int_{\alpha_{x_j}} |f_j - f_\infty + 2f_\infty|^2 \, dt \right)^{1/4} \\
&\le \epsilon^{1/4} \left(\int_{\alpha_{x_j}} (|f_j - f_\infty| + 2|f_\infty|)^2 \, dt \right)^{1/4} \\
&= \epsilon^{1/4} \left(\int_{\alpha_{x_j}} |f_j - f_\infty|^2 + 4|f_j - f_\infty| \, |f_\infty| + 4|f_\infty|^2 \, dt \right)^{1/4} \\
&\le \epsilon^{1/4} \left(\epsilon + 4c \int_{\alpha_{x_j}} |f_j - f_\infty| \, dt + 4c^2 \right)^{1/4} \\
&\le \epsilon^{1/4} \left(\epsilon + 4c \left(\int_{\alpha_{x_j}} |f_j - f_\infty|^2 \, dt \right)^{1/2} + 4c^2 \right)^{1/4} \\
&\le \epsilon^{1/4} \left(\epsilon + 4\, c\, \epsilon^{1/2} + 4c^2 \right)^{1/4}.
\end{aligned}$$

Substituting this into (41) we have

$$d_j(p,q) \le 4\delta_j + d_\infty(p,q) + \Delta z \epsilon^{1/4} \left(\epsilon + 4c\epsilon^{1/2} + 4c^2 \right)^{1/4}.$$

Now by taking limits as $j \to \infty$ we find

$$\limsup_{j \to \infty} d_j(p,q) \le d_\infty(p,q) + \Delta z \epsilon^{1/4} \left(\epsilon + 4c\epsilon^{1/2} + 4c^2 \right)^{1/4}.$$

Since this is true for all $\epsilon > 0$ the lemma follows. □

5C. ***Proof of Theorem 5.1.*** In this section we finish the proof of Theorem 5.1, which follows by the results of the last two subsections combined with Theorem 2.4.

Proof. Let $p, q \in [-\pi, \pi]^3$. Then by Lemma 4.1 we have

$$\liminf_{j\to\infty} d_j(p,q) \geq d_\infty(p,q). \tag{42}$$

By Lemmas 5.4, 5.5 or 5.6 we have

$$\limsup_{j\to\infty} d_j(p,q) \leq d_\infty(p,q).$$

So by combining with (42) we conclude

$$\lim_{j\to\infty} d_j(p,q) = d_\infty(p,q), \tag{43}$$

which gives pointwise convergence of distances.

Now by the assumption that $0 < c - \frac{1}{j} \leq f_j \leq K$ we can apply Lemma 2.3 and choose $\lambda = \max(1/\min(c/2, 1), \max(1, K)) > 0$ so that for j chosen large enough we find

$$\lambda \geq \frac{d_j(p,q)}{d_1(p,q)} \geq \frac{1}{\lambda},$$

where d_1 is the distance defined with warping factor 1.

Hence we can apply Theorem 2.4 to conclude that there exists a length metric d'_∞ and a subsequence d_{j_k} such that d_{j_k} converges uniformly to d'_∞, and GH and SWIF converges as well. By the pointwise convergence (43) we know that $d_\infty = d'_\infty$ and hence d_{j_k} must uniformly converge to d_∞. Since this is true for all the subsequences, we see that d_j uniformly converges to d_∞ and hence Gromov–Hausdorff and intrinsic flat converges as well. □

Acknowledgements

The authors would like to thank the Fields Institute and particularly Spyros Alexakis (University of Toronto), Walter Craig (McMaster University), Robert Haslhofer (University of Toronto), Spiro Karigiannis (University of Waterloo), Aaron Naber (Northwestern University), McKenzie Wang (McMaster University) for organizing the Thematic Program and the Summer School on Geometric Analysis there. It provided a wonderful place for the two of us to work and meet with new people. We'd like to thank Christian Ketterer, Chen-Yun Lin, and Raquel Perales for serving as TAs to the students attending the second author's series of talks there. Brian Allen would like to thank the United States Military Academy Department of Mathematics for funding his trip to join this team. Much of the work in this paper resulted from

discussions there as to what was needed to complete the projects the teams were working on. We wrote this paper to serve as a tool that could be applied by those teams as they meet again in the future. All graphics in this paper were drawn by Penelope Chang of Hunter College High School, NYC.

References

[Alexander and Bishop 2004] S. B. Alexander and R. L. Bishop, "Curvature bounds for warped products of metric spaces", *Geom. Funct. Anal.* **14**:6 (2004), 1143–1181. MR Zbl

[Allen 2018a] B. Allen, "IMCF and the stability of the PMT and RPI under L^2 convergence", *Ann. Henri Poincaré* **19**:4 (2018), 1283–1306. MR Zbl

[Allen 2018b] B. Allen, "Stability of the PMT and RPI for asymptotically hyperbolic manifolds foliated by IMCF", *J. Math. Phys.* **59**:8 (2018), art. id. 082501. MR Zbl

[Allen et al. 2019] B. Allen, L. Hernandez-Vazquez, D. Parise, A. Payne, and S. Wang, "Warped tori with almost non-negative scalar curvature", *Geom. Dedicata* **200** (2019), 153–171. MR Zbl

[Burago et al. 2001] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*, Graduate Studies in Math. **33**, Amer. Math. Soc., Providence, RI, 2001. MR Zbl

[Burtscher 2015] A. Y. Burtscher, "Length structures on manifolds with continuous Riemannian metrics", *New York J. Math.* **21** (2015), 273–296. MR Zbl

[Gromov 1981] M. Gromov, *Structures métriques pour les variétés riemanniennes*, Textes Math. **1**, CEDIC, Paris, 1981. MR Zbl

[Gromov 2018] M. Gromov, "Scalar curvature of manifolds with boundaries: natural questions and artificial constructions", 2018. Notes from the workshop *Emerging topics: scalar curvature and convergence* (IAS Princeton, 2018). arXiv

[Huang et al. 2017] L.-H. Huang, D. A. Lee, and C. Sormani, "Intrinsic flat stability of the positive mass theorem for graphical hypersurfaces of Euclidean space", *J. Reine Angew. Math.* **727** (2017), 269–299. MR Zbl

[Lakzian 2016] S. Lakzian, "On diameter controls and smooth convergence away from singularities", *Differential Geom. Appl.* **47** (2016), 99–129. MR Zbl

[Lakzian and Sormani 2013] S. Lakzian and C. Sormani, "Smooth convergence away from singular sets", *Comm. Anal. Geom.* **21**:1 (2013), 39–104. MR Zbl

[Lee and Sormani 2014] D. A. Lee and C. Sormani, "Stability of the positive mass theorem for rotationally symmetric Riemannian manifolds", *J. Reine Angew. Math.* **686** (2014), 187–220. MR Zbl

[Park et al. 2018] J. Park, W. Tian, and C. Wang, "A compactness theorem for rotationally symmetric Riemannian manifolds with positive scalar curvature", *Pure Appl. Math. Q.* **14**:3-4 (2018), 529–561.

[Sormani 2017] C. Sormani, "Scalar curvature and intrinsic flat convergence", pp. 288–338 in *Measure theory in non-smooth spaces*, edited by N. Gigli, De Gruyter, Warsaw, 2017. MR

[Sormani and Wenger 2011] C. Sormani and S. Wenger, "The intrinsic flat distance between Riemannian manifolds and other integral current spaces", *J. Differential Geom.* **87**:1 (2011), 117–199. MR Zbl

[Wenger 2011] S. Wenger, "Compactness for manifolds and integral currents with bounded diameter and volume", *Calc. Var. Partial Differential Equations* **40**:3-4 (2011), 423–448. MR Zbl

Received August 9, 2018. Revised February 20, 2019.

BRIAN ALLEN
UNIVERSITY OF HARTFORD
HARTFORD, CT
UNITED STATES
brianallenmath@gmail.com

CHRISTINA SORMANI
DEPARTMENT OF MATHEMATICS
CUNY GRADUATE CENTER
NEW YORK, NY
UNITED STATES
sormanic@gmail.com

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.47

EXPLICIT FORMULAE AND DISCREPANCY ESTIMATES FOR a-POINTS OF THE RIEMANN ZETA-FUNCTION

SIEGFRED BALUYOT AND STEVEN M. GONEK

For a fixed $a \neq 0$, an a-point of the Riemann zeta-function is a complex number $\rho_a = \beta_a + i\gamma_a$ such that $\zeta(\rho_a) = a$. Recently J. Steuding estimated the sum

$$\sum_{\substack{0<\gamma_a\leq T\\ \beta_a>0}} x^{\rho_a}$$

for a fixed x as $T \to \infty$, and used this to prove that the ordinates γ_a are uniformly distributed modulo 1. We provide uniform estimates for this sum when $x > 0$ and $\neq 1$, and $T > 1$. Using this, we bound the discrepancy of the sequence $\lambda\gamma_a$ when $\lambda \neq 0$. We also find explicit representations and bounds for the Dirichlet coefficients of the series $1/(\zeta(s) - a)$ and upper bounds for the abscissa of absolute convergence of this series.

1. Introduction and Results

Let $\zeta(s)$ denote the Riemann zeta-function, where $s = \sigma + it$ is a complex variable. As is usual, we shall denote zeros of the zeta-function by $\rho = \beta + i\gamma$. If a is a nonzero complex number, an *a-point* of $\zeta(s)$ is a number $\rho_a = \beta_a + i\gamma_a$ such that $\zeta(\rho_a) = a$. That is, it is a zero of $F(s) = \zeta(s) - a$. For basic results about a-points we refer the reader to [Levinson 1975; Selberg 1992; Titchmarsh 1986]. In particular, it is known that there exists a number $n_0(a)$ such that for each $n \geq n_0(a)$ there is an a-point very close to $s = -2n$, and there are at most finitely many other a-points in $\sigma \leq 0$. We call these the *trivial* a-points, and the remaining a-points *nontrivial*. Since a Dirichlet series that is not identically zero has a right half-plane free of zeros, the nontrivial a-points lie in a strip $0 < \sigma < A$, where A depends on a. It was proved in the paper of Bohr, Landau, and Littlewood [Bohr et al. 1913] that the number of these with $0 < \gamma_a \leq T$ is

$$N_a(T) = \sum_{\substack{0<\gamma_a\leq T\\ \beta_a>0}} 1 = \frac{T}{2\pi}\log\frac{T}{2\pi} - \frac{T}{2\pi} + O_a(\log T) \tag{1-1}$$

Research of the authors was partially supported by National Science Foundation grant DMS 1200582.
MSC2010: 11M06, 11M26.
Keywords: Riemann zeta-function, a-points, uniform distribution, discrepancy.

provided that $a \neq 1$; if $a = 1$ there is an additional term $-\log 2(T/2\pi)$ on the right-hand side. The corresponding formula for the number of nontrivial zeros of the zeta-function is

$$N(T) = \sum_{0<\gamma\leq T} 1 = \frac{T}{2\pi}\log\frac{T}{2\pi} - \frac{T}{2\pi} + O(\log T).$$

It was also proved in [Bohr et al. 1913] that if the Riemann hypothesis is true, the a-points cluster about the line $\sigma = \frac{1}{2}$. Much later Levinson [1975] showed that this holds unconditionally. A similar clustering result was proved for the zeros of the zeta-function by Bohr and Landau [1914]. Despite these similarities, there is a striking difference between the distribution of a-points and zeros: for each fixed σ with $\frac{1}{2} < \sigma \leq 1$ the number of a-points with $\beta_a > \sigma$ and $0 < \gamma_a \leq T$ is $\gg T$, whereas $\zeta(s)$ has only $o(T)$ zeros in this region [Titchmarsh 1986].

Landau [1912] proved the remarkable formula

$$\sum_{0<\gamma\leq T} x^{\rho} = -\Lambda(x)\frac{T}{2\pi} + O(\log T) \quad (T \to \infty),$$

where $x > 1$ is fixed. Here $\Lambda(x)$ is von Mangoldt's function defined as $\Lambda(n) = \log p$ if $n = p^k$ for some natural number k, and $\Lambda(x) = 0$ for all other real x. A formula for $0 < x < 1$ follows on replacing x by $1/x$, multiplying the resulting sum by x, and observing that $1 - \rho$ runs through the nontrivial zeros as ρ does. The two x-ranges may be combined and stated as

$$\sum_{0<\gamma\leq T} x^{\rho} = -\Big(\Lambda(x) + x\Lambda\Big(\frac{1}{x}\Big)\Big)\frac{T}{2\pi} + O(\log T) \quad (T \to \infty), \tag{1-2}$$

for any fixed positive $x \neq 1$. Recently, Steuding [2014, Theorem 6] proved an analogous formula for a-points, namely,

$$\sum_{0<\gamma_a\leq T} x^{\rho_a} = -\Big(\Lambda_a(x) + x\Lambda\Big(\frac{1}{x}\Big)\Big)\frac{T}{2\pi} + O(T^{1/2+\varepsilon}), \tag{1-3}$$

where $x \neq 1$ is fixed and positive and $\varepsilon > 0$ is arbitrarily small. When $a \neq 1$, $\Lambda_a(n)$ is defined for integers $n \geq 2$ by means of the Dirichlet series

$$-\frac{\zeta'(s)}{\zeta(s) - a} = \sum_{n=2}^{\infty}\frac{\Lambda_a(n)}{n^s}. \tag{1-4}$$

For other real x, $\Lambda_a(x) = 0$. When $a = 1$, Λ_a is defined for numbers $m2^r$ with m an odd positive integer and r any integer by means of the generalized Dirichlet

series

$$-\frac{\zeta'(s)}{\zeta(s)-1}=\sum_{\substack{m=1\\ \text{odd}}}^{\infty}\sum_{r=-\infty}^{\infty}\frac{\Lambda_1(2^r m)}{(2^r m)^s}.$$

Here too $\Lambda_1(x)=0$ for other real x.

The implied constants in (1-2) and (1-3) are highly dependent on x. For example, in the case of (1-2), Gonek [1985; 1993] proved that when $x, T>1$

$$\sum_{0<\gamma\le T}x^{\rho}=-\frac{T}{2\pi}\Lambda(x)+O(x\log(2xT)\log\log(3x))+O\Big(\log x\min\Big(T,\frac{x}{\langle x\rangle}\Big)\Big)$$
$$+O\Big(\log(2T)\min\Big(T,\frac{1}{\log x}\Big)\Big),$$

where $\langle x\rangle$ denotes the distance from x to the nearest prime power other than x itself, and the implied constants in the O-terms are absolute. An immediate corollary of this is that for $x, T>1$ we have

$$\sum_{0<\gamma\le T}x^{-\rho}=-\frac{T}{2\pi x}\Lambda(x)+O(\log(2xT)\log\log(3x))+O\Big(\log x\min\Big(\frac{T}{x},\frac{1}{\langle x\rangle}\Big)\Big)$$
$$+O\Big(\log(2T)\min\Big(\frac{T}{x},\frac{1}{x\log x}\Big)\Big).$$

Our first aim here is to prove analogues of these formulae for (1-3).

In stating our results it will be convenient to write

$$\beta_a^*=\sup_{\rho_a}\beta_a \tag{1-5}$$

and

$$B=\beta_a^*+\varepsilon,$$

where $\varepsilon>0$ is arbitrary. Thus, the value of B may be different at different occurrences. As was mentioned above, there is a number A such that all $\beta_a<A$, so β_a^* is finite. Furthermore, (see Theorem 11.6(C) of [Titchmarsh 1986]) we know that for every $\delta>0$ the equation $\zeta(s)=a$ has solutions in the strip $1<\sigma<1+\delta$. Thus $\beta_a^*>1$.

Theorem 1.1. *Suppose $a\neq 0, 1$ is a fixed complex number and let $x, T>1$. Then*

$$\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}}x^{\rho_a}=-\frac{T}{2\pi}\Lambda_a(x)+O\Big(x^B\Big(1+\min\Big\{T,\frac{x}{\langle x\rangle}\Big\}\Big)\Big)$$
$$+O\Big(x^{B+1}\log T\Big(1+\frac{1}{\log x}\Big)\Big)+O\Big(\frac{\log T}{x^2}\Big(1+\min\Big\{T,\frac{1}{\log x}\Big\}\Big)\Big).$$

The implied constants depend only on a and the value of ε in the definition of B.

To estimate

$$\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}} x^{\rho_a} \quad \text{when } 0<x<1,$$

we consider

$$\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}} x^{-\rho_a} \quad \text{with } x>1.$$

In this case we do not need to exclude $a=1$.

Theorem 1.2. *Let $a\neq 0$ and $0<\theta<1$ be fixed. If $T>1$ and $1<x\le T^{\theta}$, then*

$$\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}} x^{-\rho_a} = -\frac{T}{2\pi x}\Lambda(x)+O\Big(\frac{\log T}{\log x}\Big)+O\Big(\log(2x)\min\Big\{\frac{T}{x},\frac{1}{\langle x\rangle}\Big\}\Big)+O(\log^4 T).$$

It would be interesting to have a version of Theorem 1.1 when $a=1$ also. This looks possible but rather complicated and is not needed for the applications below.

Steuding [2014] used (1-3) to prove the interesting result that the fractional parts of the sequence $\{\lambda\gamma_a\}_{\gamma_a>0}$ are uniformly distributed modulo 1, where λ is any fixed nonzero real number.[1] Our uniform versions of (1-3) allow us to prove a discrepancy estimate for this sequence.

Theorem 1.3. *Let $a\neq 0$ and let $\lambda\neq 0$ be a fixed real number. Then for T sufficiently large we have*

$$\sup_{0\le\alpha\le 1}\Big|\frac{1}{N_a(T)}\Big(\sum_{\substack{0<\gamma_a\le T,\ \beta_a>0\\ \{\lambda\gamma_a\}\le\alpha}} 1\Big)-\alpha\Big| \ll \frac{1}{\log\log T}, \tag{1-6}$$

where $\{x\}$ denotes the fractional part of the real number x.

The analogous problem for the zeros (i.e., the "case $a=0$" of Theorem 1.3) has been studied extensively. The interested reader is referred to the survey [Steuding 2014] for an informative discussion of this problem and related results.

As another application of Theorem 1.2 we prove

Theorem 1.4. *Let*

$$A(s)=\sum_{n\le N} a(n)n^{-s},$$

[1]The statement of Theorem 6 in [Steuding 2014] is incorrect when $a=1$ because in that case $-\zeta'(s)/(\zeta(s)-1)$ cannot be expressed as an ordinary Dirichlet series.

where the $a(n)$ are complex numbers such that $|a(n)| \ll n^{\varepsilon}$ and $N = T^{\theta}$ with $T \geq 2$ and $0 < \theta < 1$ fixed. Then for $a \neq 0$ we have

$$\sum_{\substack{0<\gamma_a\leq T\\ \beta_a>0}} A(\rho_a) = \frac{T}{2\pi}\left(a(1)\log T - \sum_{2\leq n\leq N}\frac{a(n)\Lambda(n)}{n}\right) + O(T).$$

Specializing the Dirichlet polynomial $A(s)$ leads to the following formulae.

Corollary 1.5. *Let*

$$M(s) = \sum_{n\leq N}\frac{\mu(n)}{n^s} \quad \text{and} \quad P(s) = \sum_{n\leq N}\frac{1}{n^s}.$$

If $N = T^{\theta}$ with $0 < \theta < 1$ fixed and $a \neq 0$, then

$$\sum_{\substack{0<\gamma_a\leq T\\ \beta_a>0}} M(\rho_a) = (1+\theta)\frac{T}{2\pi}\log T + O(T) \tag{1-7}$$

and

$$\sum_{\substack{0<\gamma_a\leq T\\ \beta_a>0}} P(\rho_a) = (1-\theta)\frac{T}{2\pi}\log T + O(T). \tag{1-8}$$

These results seemed counterintuitive to us at first. To the extent that one expects $M(s)$ to approximate $1/\zeta(s)$ and $P(s)$ to approximate $\zeta(s)$ on average, one might expect the first sum to be large and the second small when $|a|$ is small, and expect the reverse to be true when $|a|$ is large. However, from the corollary we see that the first sum is always larger than the second. The explanation seems to be that many a-points are quite close to zeros of $\zeta(s)$. In fact, the same argument as in the proof of the corollary shows that (1-7) and (1-8) hold with the ρ_a's replaced by ρ's.

Theorems 1.1 and 1.2 are proved by calculating the integrals

$$\frac{1}{2\pi i}\int_{\mathscr{R}} x^{\pm s}\frac{\zeta'(s)}{\zeta(s)-a}ds$$

over an appropriate rectangle $\mathscr{R}$. The size of the coefficients of the Dirichlet series for $1/(\zeta(s)-a)$, and its abscissae of convergence and absolute convergence enter into this analysis, so we shall also prove the following results. Although we do not require as much detail as the next two theorems provide, we record them in the hope that they may prove useful to others.

Theorem 1.6. *For $a \neq 0, 1$ the coefficients of the Dirichlet series*

$$\frac{1}{\zeta(s)-a} = \sum_{n=1}^{\infty}\frac{b_a(n)}{n^s}$$

are given by

$$b_a(n) = \begin{cases} -\sum_{k=0}^{\infty} a^{-k-1} d_k(n) & \text{if } |a| > 1, \\ \sum_{k=1}^{\infty} a^{k-1} d_{-k}(n) & \text{if } 0 < |a| < 1, \\ -\sum_{k=0}^{\infty} (a-1)^{-k-1} e_k(n) & \text{if } |a| = 1, \text{ but } a \neq 1. \end{cases}$$

Here $d_l(n)$ is the n-th Dirichlet coefficient of $\zeta(s)^l$ and $e_l(n)$ is the n-th Dirichlet coefficient of $(\zeta(s)-1)^l$. When $a = 1$ the series is the generalized Dirichlet series

$$\frac{1}{\zeta(s)-1} = \sum_{\substack{m=1 \\ \text{odd}}}^{\infty} \sum_{r=-\infty}^{\infty} \frac{b_1(m2^r)}{(m2^r)^s}$$

with coefficients

$$b_1(m2^r) = \sum_{\substack{l-k-1=r \\ k,l \geq 0}} (-1)^k f_k(2^l m),$$

where $f_k(n)$ is the n-th Dirichlet coefficient of $(\zeta(s) - 1 - 2^{-s})^k$.

Theorem 1.7. *Let $a \neq 0$. Define $\sigma^* > 1$ to be the unique solution to the equation*

$$\begin{cases} \zeta(\sigma) = |a| & \text{if } |a| > 1, \\ \zeta(\sigma) = 1 + |1-a| & \text{if } |a| = 1,\ a \neq 1, \\ \dfrac{\zeta(2\sigma)}{\zeta(\sigma)} = |a| & \text{if } |a| < 1, \\ \zeta(\sigma) = 1 + 2^{1-\sigma} & \text{if } a = 1. \end{cases}$$

Then the abscissa of absolute convergence $\bar{\sigma}$ of the series for $1/(\zeta(s)-a)$ satisfies

$$\bar{\sigma} \leq \sigma^*.$$

Remark. When $a = 1$, $\sigma^* \approx 2.4241$.

Theorem 1.8. *Let $a \neq 0, 1$ and let σ_0 be the abscissa of convergence of the Dirichlet series*

$$\frac{1}{\zeta(s)-a} = \sum_{n=1}^{\infty} \frac{b_a(n)}{n^s}.$$

Then $\sigma_0 = \beta_a^*$, with β_a^* as in (1-5)*, and*

$$1 < \sigma_0 \leq \bar{\sigma} \leq \sigma^*.$$

Moreover, for every $\varepsilon > 0$

$$b_a(n) \ll n^{\beta_a^* + \varepsilon}.$$

This bound is sharp in the sense that

$$|b_a(n)| > n^{\beta_a^* - \varepsilon}$$

for infinitely many n.

2. Proof of Theorem 1.1

In the following proof we shall appeal to Theorem 1.8 though it is proved later.

The functional equation for $\zeta(s)$ is

$$\zeta(s) = \chi(s)\zeta(1-s), \tag{2-1}$$

where, by Stirling's formula,

$$\chi(s) = \Big(\frac{t}{2\pi}\Big)^{1/2-s} e^{i\pi/4+it}\Big(1 + O\Big(\frac{1}{t}\Big)\Big) \tag{2-2}$$

as $t \to \infty$ in any fixed vertical strip. From (3.11.8) of [Titchmarsh 1986], we have

$$|\zeta(s)| \gg \frac{1}{\log t}$$

as $t \to \infty$ when $\sigma \geq 1 - A/\log t$ and, in particular, when $\sigma \geq 1$. Thus, from (2-1) and (2-2) we have

$$|\zeta(s)| \gg \frac{t^{1/2-\sigma}}{\log t} \tag{2-3}$$

as $t \to \infty$ in any fixed vertical strip with $\sigma \leq 0$. We may therefore choose a number $T_0 \geq 2$ such that $|\zeta(s)| > |a|$ for $\sigma \leq 0$ and $t \geq T_0$, and also so that no γ_a equals T_0. With this T_0, and any $T > T_0$ with $T \neq \gamma_a$ for any γ_a, consider the contour integral

$$\begin{aligned} I &= \frac{1}{2\pi i}\Big(\int_{B+1+iT_0}^{B+1+iT} + \int_{B+1+iT}^{-2+iT} + \int_{-2+iT}^{-2+iT_0} + \int_{-2+iT_0}^{B+1+iT_0}\Big)\Big(\frac{\zeta'(s)}{\zeta(s)-a}\Big)x^s\,ds \\ &= I_1 + I_2 + I_3 + I_4, \end{aligned}$$

say. By the calculus of residues

$$I = \sum_{\substack{T_0<\gamma_a<T \\ \beta_a>0}} x^{\rho_a}.$$

To prove the theorem we estimate I_1 through I_4.

To estimate I_1 we use the Dirichlet series expansion (1-4), which by Theorem 1.8 is absolutely convergent for $\sigma = B+1$, and integrate term-by-term. This leads to

$$\begin{aligned} I_1 &= -\sum_{n=2}^{\infty} \Lambda_a(n)\Big(\frac{x}{n}\Big)^{B+1}\Big(\frac{1}{2\pi}\int_{T_0}^{T}\Big(\frac{x}{n}\Big)^{it}dt\Big) \\ &= -\frac{T-T_0}{2\pi}\Lambda_a(x) + O\Big(\sum_{\substack{n=2 \\ n\neq x}}^{\infty} \frac{|\Lambda_a(n)|}{n^{B+1}}x^{B+1}\min\Big\{T, \frac{1}{|\log(x/n)|}\Big\}\Big). \end{aligned} \tag{2-4}$$

To estimate the sum in the error term note that $|\log(x/n)| \gg 1$ for $n \le x/2$ or $n \ge 2x$. Thus the part of the sum with $n \le x/2$ or $n \ge 2x$ is

$$\ll \sum_{n=1}^{\infty} \frac{|\Lambda_a(n)|}{n^{B+1}} x^{B+1} \ll x^{B+1}. \tag{2-5}$$

The part with $x/2 < n < x$ is

$$\begin{aligned} &\ll \sum_{x/2<n<x} |\Lambda_a(n)| \min\Big\{T, \frac{1}{\log(x/n)}\Big\} \\ &= \sum_{x/2<n<N} |\Lambda_a(n)| \min\Big\{T, \frac{1}{\log(x/n)}\Big\} + |\Lambda_a(N)| \min\Big\{T, \frac{1}{\log(x/N)}\Big\}, \end{aligned}$$

where N is the largest integer less than x. By Theorem 1.8, we have $\Lambda_a(n) \ll_\varepsilon n^B$. Thus, since

$$\log \frac{x}{n} = -\log\Big(1 - \frac{x-n}{x}\Big) > \frac{N-n}{x},$$

we see that

$$\sum_{x/2<n<N} |\Lambda_a(n)| \min\Big\{T, \frac{1}{\log x/n}\Big\} \le x \sum_{x/2<n<N} \frac{|\Lambda_a(n)|}{N-n} \ll_\varepsilon x N^B \log x \ll_\varepsilon x^{B+1}.$$

On the other hand, we have

$$|\Lambda_a(N)| \min\Big\{T, \frac{1}{\log x/N}\Big\} \ll_\varepsilon N^B \min\Big\{T, \frac{x}{x-N}\Big\} \ll x^B \min\Big\{T, \frac{x}{\langle x\rangle}\Big\}.$$

Hence the part with $x/2 < n < x$ is

$$\ll_\varepsilon x^{B+1} + x^B \min\Big\{T, \frac{x}{\langle x\rangle}\Big\}.$$

A similar argument gives the same estimate for the part with $x < n < 2x$. Using this and (2-5) in (2-4), we obtain

$$I_1 = -\frac{T}{2\pi}\Lambda_a(x) + O_\varepsilon\Big(x^{B+1} + x^B \min\Big\{T, \frac{x}{\langle x\rangle}\Big\}\Big). \tag{2-6}$$

To estimate I_2, we require the following lemma.

Lemma 2.1. *There is a positive number R_a depending only on a such that for $R \ge R_a$ we have*

$$\frac{\zeta'(s)}{\zeta(s)-a} = \sum_{|\rho_a - s|<R} \frac{1}{s-\rho_a} + O_R(\log t)$$

uniformly for $-2 \le \sigma \le R-2$ and large t.

Proof. Let $f(s)=\zeta(s)-a$. If $r_a > \beta_a^*$ is large enough, then for $\sigma_0 \geq r_a$ we will have $|f(\sigma_0+it)| \gg_{\sigma_0} 1$ for all large t. We will show how to determine such an r_a later. We apply Lemma α of §3.9 in [Titchmarsh 1986] with $f(s)=\zeta(s)-a$, $s_0=\sigma_0+iT$, $r=4(\sigma_0+2)$, and T large. By the Phragmén–Lindelöf theorem applied to $\zeta(s)$ (see, for example, Chapter 5 of [Titchmarsh 1932]), we have $f(s)=O_r(T^A)$ for some constant A uniformly for $|s-s_0|\leq r$. Thus

$$\left|\frac{f(s)}{f(s_0)}\right| \ll_r T^A$$

uniformly for $|s-s_0|\leq r$. It now follows from [Titchmarsh 1986, Lemma α] that

$$\frac{\zeta'(s)}{\zeta(s)-a} = \sum_{|\rho_a - s|\leq r/4} \frac{1}{s-\rho_a} + O_r(\log T)$$

for $|s-s_0|\leq r/4$. If $s=\sigma+iT$ and $-2\leq\sigma\leq 2\sigma_0+2$, then $|s-s_0|\leq r/4$ because $r=4(\sigma_0+2)$. This proves the lemma with $R_a=4(r_a+2)$ and $R=r/4$.

We now show how to choose an r_a such that if $\sigma_0\geq r_a$ then $|f(\sigma_0+it)|\gg_{\sigma_0} 1$ for all large t. If $a\neq 1$, then $|1-a|\neq 0$. Hence, since $\lim_{\sigma\to 1}\zeta(\sigma)=1$, we may choose a number σ_1 so large that $|1-a|>\zeta(\sigma)-1$ for $\sigma\geq\sigma_1$. If $a=1$, we choose $\sigma_1=4$. In that case $\sigma\geq\sigma_1$ implies

$$\sum_{n=3}^{\infty}\frac{1}{n^\sigma} \leq \int_2^\infty \frac{1}{u^\sigma}\,d\sigma = \frac{2^{1-\sigma}}{\sigma-1} \leq \frac{2}{3}\cdot\frac{1}{2^\sigma},$$

which in turn implies that

$$|\zeta(s)-1| = \left|2^{-s}+\sum_{n=3}^{\infty} n^{-s}\right| \geq 2^{-\sigma}/3.$$

We now set $r_a=\max\{\sigma_1, \beta_a^*+1\}$. It then follows that if $a\neq 1$ and $\sigma_0\geq r_a$, then

$$|f(\sigma_0+it)| = |\zeta(\sigma_0+it)-1+(1-a)| \geq |1-a|-\zeta(\sigma_0)+1>0.$$

On the other hand, if $a=1$ and $\sigma_0\geq r_a$, then

$$|f(\sigma_0+it)| = |\zeta(\sigma_0+it)-1| \geq 3^{-1}2^{-\sigma_0}>0.$$

This completes the proof. □

By Lemma 2.1,

$$I_2 = \sum_{|\rho_a-s|<R_a} \frac{1}{2\pi i}\int_{B+1+iT}^{-2+iT} \frac{x^s}{s-\rho_a}\,ds + O\left(\log T\int_{-2}^{B+1} x^\sigma\,d\sigma\right).$$

The error term is

$$\ll \log T \, \frac{x^{B+1}}{\log x}.$$

To estimate the sum, note that by Cauchy's integral theorem we may replace the line segment of integration in each term by the semicircle above or below the segment depending on whether ρ_a lies below or above that segment. Thus, the sum is

$$\ll \sum_{|\rho_a - s| < R} x^{B+1}.$$

By (1-1) the number of terms in the sum is $O(\log T)$. Thus,

$$I_2 \ll x^{B+1} \log T \Big(1 + \frac{1}{\log x}\Big). \tag{2-7}$$

To estimate I_3 note that by our choice of T_0, if $\sigma = -2$ and $t \geq T_0$, then

$$\frac{1}{\zeta(s) - a} = \frac{1}{\zeta(s)} \Big(\frac{1}{1 - a/\zeta(s)}\Big) = \frac{1}{\zeta(s)} \sum_{k=0}^{\infty} \Big(\frac{a}{\zeta(s)}\Big)^k.$$

Thus

$$I_3 = \frac{1}{2\pi i} \int_{-2+iT}^{-2+iT_0} x^s \frac{\zeta'(s)}{\zeta(s)}\, ds + \frac{1}{2\pi i} \int_{-2+iT}^{-2+iT_0} x^s \frac{\zeta'(s)}{\zeta(s)} \sum_{k=1}^{\infty} \Big(\frac{a}{\zeta(s)}\Big)^k ds. \tag{2-8}$$

From the logarithmic derivative of the functional equation for $\zeta(s)$ (for example, see [Gonek 1993] or [Davenport 1980, pp. 73, 80, 81]) we have

$$-\frac{\zeta'(s)}{\zeta(s)} = \frac{\zeta'(1-s)}{\zeta(1-s)} + \log \frac{t}{2\pi} + O\Big(\frac{1}{t}\Big) \tag{2-9}$$

in any half-strip $A_1 \leq \sigma \leq A_2$, $t \geq 1$ that does not contain zeros of $\zeta(s)$. Thus, the first integral on the right-hand side of (2-8) equals

$$\frac{x^{-2}}{2\pi} \int_{T_0}^{T} x^{it} \frac{\zeta'}{\zeta}(3 - it)\, dt + \frac{x^{-2}}{2\pi} \int_{T_0}^{T} x^{it} \log \frac{t}{2\pi}\, dt + O(x^{-2} \log T).$$

We insert the Dirichlet series for ζ'/ζ into the first integral here and integrate term-by-term, and in the second we integrate by parts. In this way we find that

$$\frac{1}{2\pi i} \int_{-2+iT}^{-2+iT_0} x^s \frac{\zeta'(s)}{\zeta(s)}\, ds \ll \frac{\log T}{x^2} + \min\Big\{\frac{T \log T}{x^2}, \frac{\log T}{x^2 \log x}\Big\}. \tag{2-10}$$

To estimate the second integral on the right-hand side of (2-8), note that by(2-3) we have

$$\zeta(-2 + it) \gg \frac{t^{5/2}}{\log t}$$

for $t \geq T_0$. Also, by (2-9), we have

$$\frac{\zeta'(-2+it)}{\zeta(-2+it)} \ll \log t$$

for $t \geq T_0$. Hence

$$\frac{1}{2\pi i}\int_{-2+iT}^{-2+iT_0} x^s \frac{\zeta'(s)}{\zeta(s)} \sum_{k=1}^{\infty}\Big(\frac{a}{\zeta(s)}\Big)^k ds \ll x^{-2}\int_{T_0}^{T} \frac{\log^2 t}{t^{5/2}}\, dt \ll x^{-2}.$$

From this and (2-10) we obtain

$$I_3 \ll \frac{\log T}{x^2} + \min\Big\{\frac{T\log T}{x^2}, \frac{\log T}{x^2\log x}\Big\}. \tag{2-11}$$

Finally, $\zeta'(s)/(\zeta(s)-a)$ is bounded on $[-2+iT_0, B+1+iT_0]$, so

$$I_4 \ll x^{B+1}.$$

Combining this, (2-6), (2-7), and (2-11), we find that for $T \geq T_0$

$$\begin{aligned}\sum_{\substack{T_0<\gamma_a<T\\ \beta_a>0}} x^{\rho_a} = -\frac{T}{2\pi}\Lambda_a(x) &+ O_\varepsilon\Big(x^{B+1} + x^B \min\Big\{T, \frac{x}{\langle x\rangle}\Big\}\Big)\\ &+ O\Big(x^{B+1}\log T\Big(1+\frac{1}{\log x}\Big)\Big)\\ &+ O\Big(\frac{\log T}{x^2} + \frac{\log T}{x^2}\min\Big\{T, \frac{1}{\log x}\Big\}\Big).\end{aligned} \tag{2-12}$$

Recall that we have assumed $T \neq \gamma_a$ for any γ_a. To remove this assumption, observe that by (1-1), changing T by a bounded amount in (2-12) changes the value of the sum on the left-hand side by at most $O(x^B \log T)$. This is clearly no more than the resulting change on the right-hand side.

As we mentioned in the first paragraph of Section 1 all the nontrivial a-points lie in a strip of the form $0 < \sigma < A$. There are at most a finite number of these with $0 < \gamma_a \leq T_0$, hence

$$\sum_{\substack{0<\gamma_a\leq Y\\ \beta_a>0}} x^{\rho_a} \ll x^B \tag{2-13}$$

uniformly for $1 < Y \leq T_0$. Taking $Y = T_0$ and combining this with (2-12), we see that we may extend the sum on the left-hand side of (2-12) to run over all ρ_a with $0 < \gamma_a \leq T$ and $\beta_a > 0$. The resulting formula holds for $T \geq T_0 \geq 2$. To see that it also holds when T is between 1 and T_0, note that (2-13) holds with $Y = T$, and the right-hand side is bounded by the second error term on the right-hand side of (2-12). This completes the proof of the Theorem 1.1.

3. Proof of Theorem 1.2

Let $a \neq 0$ and suppose that $x > 1$. As in the proof of Theorem 1.1 (see below (2-3)), we can choose a $T_0 \geq 2$ such that $|\zeta(s)| > |a|$ for $\sigma \leq 0$, $t \geq T_0$, and such that no γ_a equals T_0. We also choose a $T > T_0$ which is not equal to any γ_a. With σ^* as in Theorem 1.7, we see by the calculus of residues that

$$\begin{aligned}\sum_{\substack{T_0<\gamma_a<T\\ \beta_a>0}} x^{-\rho_a} &= \frac{1}{2\pi i}\Bigg(\int_{\sigma^*+1+iT_0}^{\sigma^*+1+iT} + \int_{\sigma^*+1+iT}^{-1/\log(3x)+iT} \\ &\qquad + \int_{-1/\log(3x)+iT}^{-1/\log(3x)+iT_0} + \int_{-1/\log(3x)+iT_0}^{\sigma^*+1+iT_0}\Bigg)\left(\frac{\zeta'(s)}{\zeta(s)-a}\right)x^{-s}\,ds \\ &= I_1 + I_2 + I_3 + I_4,\end{aligned}$$

say.

To estimate I_1 we first assume $a \neq 1$. Using the Dirichlet series expansion (1-4) and integrating term-by-term, we obtain

$$\begin{aligned}I_1 &= -\sum_{n=2}^{\infty}\Lambda_a(n)\Big(\frac{1}{nx}\Big)^{\sigma^*+1}\left(\frac{1}{2\pi}\int_{T_0}^{T}\Big(\frac{1}{nx}\Big)^{it}dt\right) \\ &\ll x^{-\sigma^*-1}\sum_{n=2}^{\infty}\frac{|\Lambda_a(n)|}{n^{\sigma^*+1}\log(nx)} < \frac{1}{x^{\sigma^*+1}\log x}\sum_{n=2}^{\infty}\frac{|\Lambda_a(n)|}{n^{\sigma^*+1}} \ll \frac{1}{x^{\sigma^*+1}\log x}.\end{aligned}$$

Now assume $a = 1$. By (6-4) below we see that

$$\begin{aligned}I_1 &= \int_{\sigma^*+1+iT_0}^{\sigma^*+1+iT}\frac{\zeta'(s)}{\zeta(s)-1}x^{-s}\,ds \\ &= -\int_{\sigma^*+1+iT_0}^{\sigma^*+1+iT}\sum_{\nu=2}^{\infty}\frac{\log\nu}{\nu^s}\sum_{k=0}^{\infty}(-1)^k 2^{(k+1)s}\sum_{n=3^k}^{\infty}\frac{f_k(n)}{n^s}x^{-s}\,ds.\end{aligned}$$

Note that by Theorem 1.7 the double series over k and n converges absolutely when $\sigma = \sigma^* + 1$. Hence

$$\begin{aligned}I_1 &= -\sum_{\nu=2}^{\infty}\sum_{k=0}^{\infty}\sum_{n=3^k}^{\infty}(\log\nu)(-1)^k f_k(n)\int_{\sigma^*+1+iT_0}^{\sigma^*+1+iT}\Big(\frac{2^{k+1}}{xn\nu}\Big)^s ds \\ &\ll \sum_{\nu=2}^{\infty}\sum_{k=0}^{\infty}\sum_{n=3^k}^{\infty}\frac{(\log\nu)f_k(n)}{\log(xn\nu/2^{k+1})}\Big(\frac{2^{k+1}}{xn\nu}\Big)^{\sigma^*+1}.\end{aligned}$$

This is absolutely convergent because

$$\log\Big(\frac{xn\nu}{2^{k+1}}\Big) \geq \log\Big(\frac{3^k\cdot 2}{2^{k+1}}\Big) = k\log\frac{3}{2}$$

for $k \geq 1$, while

$$\log\Big(\frac{xn\nu}{2}\Big) \geq \log\Big(\frac{xn\cdot 2}{2}\Big) = \log xn \geq \log x > 0$$

for $k = 0$. Thus

$$I_1 \ll \frac{1}{x^{\sigma^*+1}\log x}, \tag{3-1}$$

which is the same as our estimate when $a \neq 1$.

To estimate I_2 we use Lemma 2.1 to write

$$I_2 = \sum_{|\rho_a - s| < R} \frac{1}{2\pi i} \int_{\sigma^*+1+iT}^{-1/\log(3x)+iT} \frac{x^{-s}}{s-\rho_a}\, ds + O\Big(\log T \int_{-1/\log(3x)}^{\sigma^*+1} x^{-\sigma}\, d\sigma\Big).$$

The error term is

$$\ll \log T\, \frac{x^{1/\log(3x)}}{\log x} \ll \frac{\log T}{\log x}.$$

To bound the sum, note that by Cauchy's integral theorem we may replace the path of integration in each term by the semicircle above or below the path depending on whether ρ_a lies below or above it. In this way we see that the sum is

$$\ll \sum_{|\rho_a - s| < R} x^{1/\log(3x)} \ll \sum_{|\rho_a - s| < R} 1 \ll \log T$$

by (1-1). Thus

$$I_2 \ll \log T\Big(1 + \frac{1}{\log x}\Big). \tag{3-2}$$

Next we come to I_3. Since $|\zeta(s)| > |a|$ when $\sigma \leq 0$ and $t \geq T_0$, we have

$$\frac{1}{\zeta(s)-a} = \frac{1}{\zeta(s)}\Big(\frac{1}{1-a/\zeta(s)}\Big) = \frac{1}{\zeta(s)} \sum_{k=0}^{\infty} \Big(\frac{a}{\zeta(s)}\Big)^k.$$

Hence

$$\begin{aligned} I_3 = {} & \frac{1}{2\pi i} \int_{-1/\log(3x)+iT}^{-1/\log(3x)+iT_0} x^{-s} \frac{\zeta'(s)}{\zeta(s)}\, ds \\ & + \frac{1}{2\pi i} \int_{-1/\log(3x)+iT}^{-1/\log(3x)+iT_0} x^{-s} \frac{\zeta'(s)}{\zeta(s)} \sum_{k=1}^{\infty} \Big(\frac{a}{\zeta(s)}\Big)^k\, ds \\ = {} & I_{31} + I_{32}, \end{aligned}$$

say.

We first consider I_{32}. By (3.11.7) of [Titchmarsh 1986] and (2-9) we have

$$\frac{\zeta'(s)}{\zeta(s)} \ll \log t \tag{3-3}$$

for $\sigma \le 0$ bounded and $t \ge T_0$. Using this and (2-3), we see that the terms in I_{32} with $k > 1$ contribute at most

$$\ll a^2 \int_{T_0}^{T} \frac{\log^3 t}{t^{1+2/\log(3x)}}\, dt \ll \log^4 T. \tag{3-4}$$

By integration by parts, the term with $k = 1$ is

$$-\frac{a}{2\pi i}\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT} x^{-s}\frac{\zeta'(s)}{\zeta^2(s)}\, ds$$
$$= a\frac{x^{-s}}{2\pi i\,\zeta(s)}\Big|_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT} + a\frac{\log x}{2\pi i}\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT}\frac{x^{-s}}{\zeta(s)}\, ds.$$

By (2-3) the first term on the right-hand side is $\ll T_0^{-1/2}\log T_0 \ll 1$. Hence,

$$I_{3,2} = a\frac{\log x}{2\pi i}\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT}\frac{x^{-s}}{\zeta(s)}\, ds + O(\log^4 T). \tag{3-5}$$

Using the functional equation (2-1) in the integral and switching the order of summation and integration (by absolute convergence), we see that

$$I_{3,2} = a\frac{\log x}{2\pi i}\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT}\frac{x^{-s}}{\chi(s)}\left(\sum_{n=1}^{\infty}\frac{\mu(n)}{n^{1-s}}\right) ds + O(\log^4 T)$$
$$= a\log x\sum_{n=1}^{\infty}\frac{\mu(n)}{n}\left(\frac{1}{2\pi i}\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT}\left(\frac{x}{n}\right)^{-s}\frac{1}{\chi(s)}\, ds\right) + O(\log^4 T).$$

By (2-2), we next obtain

$$I_{3,2} = \frac{ae^{-i\pi/4}}{2\pi}x^{1/\log 3x}\log x$$
$$\times\sum_{n=1}^{\infty}\frac{\mu(n)}{n^{1+1/\log(3x)}}\left(\int_{T_0}^{T}\left(\frac{t}{2\pi}\right)^{-1/2-1/\log 3x}\exp\left(it\log\frac{tn}{2\pi ex}\right)\left(1+O\left(\frac{1}{t}\right)\right)dt\right)$$
$$+O(\log^4 T).$$

The O-term inside the integral contributes

$$\ll \log 3x\sum_{n=1}^{\infty}\frac{1}{n^{1+1/\log(3x)}}\left(\int_{T_0}^{T} t^{-3/2}\, dt\right) \ll T_0^{-1/2}\log^2(3x) \ll \log^2 T.$$

Thus

$$I_{3,2} = \frac{ae^{-i\pi/4}}{2\pi}x^{1/\log 3x}\log x$$
$$\times\sum_{n=1}^{\infty}\frac{\mu(n)}{n^{1+1/\log(3x)}}\left(\int_{T_0}^{T}\left(\frac{t}{2\pi}\right)^{-1/2-1/\log 3x}\exp\left(it\log\frac{tn}{2\pi ex}\right)dt\right)+O(\log^4 T).$$

We next split the interval of integration into dyadic intervals $I_k = (T/2^{k+1}, T/2^k]$ with $k = 0, 1, 2, \dots, K = [(\log(T/T_0)/\log 2)] - 1$, plus the possible additional interval $I_{K+1} = [T_0, T/2^{K+1}] \subseteq [T_0, 2T_0]$. We then have

$$I_{3,2} = \frac{ae^{-i\pi/4}}{2\pi} x^{1/\log 3x} \log x \sum_{n=1}^{\infty} \frac{\mu(n)}{n^{1+1/\log(3x)}} \left(\sum_{k=0}^{K+1} \mathcal{J}_k(n) \right) + O(\log^4 T) \tag{3-6}$$

$$= \frac{ae^{-i\pi/4}}{2\pi} x^{1/\log 3x} \log x \sum_{k=0}^{K+1} \left(\sum_{n=1}^{\infty} \frac{\mu(n)}{n^{1+1/\log(3x)}} \mathcal{J}_k(n) \right) + O(\log^4 T),$$

where

$$\mathcal{J}_k(n) = \int_{I_k} \left(\frac{t}{2\pi} \right)^{-1/2-1/\log 3x} \exp\left(it \log \frac{tn}{2\pi ex} \right) dt. \tag{3-7}$$

To estimate this we apply the following minor modification of a lemma in [Gonek 1984].

Lemma 3.1. *For large A and B with $A < r \le B \le 2A$,*

$$\int_A^B \exp\left(it \log\left(\frac{t}{re} \right) \right) \left(\frac{t}{2\pi} \right)^{a-1/2} dt = (2\pi)^{1-a} r^a e^{-ir+\pi i/4} + O(E(r, A, B)),$$

where a is bounded and where

$$E(r, A, B) = A^{a-1/2} + \frac{A^{a+1/2}}{|A - r| + A^{1/2}} + \frac{B^{a+1/2}}{|B - r| + B^{1/2}}. \tag{3-8}$$

For $r \le A$ or $r > B$,

$$\int_A^B \exp\left(it \log\left(\frac{t}{re} \right) \right) \left(\frac{t}{2\pi} \right)^{a-1/2} dt = O(E(r, A, B)). \tag{3-9}$$

For us the cruder bound $E(r, A, B) \ll A^a$ suffices. Assuming that T_0 is sufficiently large (as we may) we then find that for $k = 0, \dots, K$,

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{\mu(n)}{n^{1+1/\log(3x)}} \mathcal{J}_k(n) &\ll \sum_{\pi x 2^{k+1}/T \le n < \pi x 2^{k+2}/T} \frac{1}{n} + \sum_{n=1}^{\infty} \frac{1}{n^{1+1/\log(3x)}} E\left(\frac{2\pi x}{n}, \frac{T}{2^{k+1}}, \frac{T}{2^k} \right) \\ &\ll \sum_{n \le x} \frac{1}{n} + \left(\frac{T}{2^k} \right)^{-1/\log(3x)} \sum_{n=1}^{\infty} \frac{1}{n^{1+1/\log(3x)}} \\ &\ll \log(3x) \left(\left(\frac{T}{2^k} \right)^{-1/\log(3x)} + 1 \right). \end{aligned} \tag{3-10}$$

We similarly find that

$$\sum_{n=1}^{\infty}\frac{\mu(n)}{n^{1+1/\log(3x)}}\mathcal{J}_{K+1}(n)\ll\log(3x).$$

Inserting these estimates in (3-6) and summing, we find that

$$\begin{aligned}I_{3,2}&\ll\log^2(3x)\sum_{k=0}^{K+1}\left(\left(\frac{2^k}{T}\right)^{1/\log(3x)}+1\right)+\log^4 T\\&\ll\log^3(3x)\left(\left(\frac{2^K}{T}\right)^{1/\log(3x)}+K\right)+\log^4 T\\&\ll\log^3(3x)\left(\left(\frac{1}{T_0}\right)^{1/\log(3x)}+\log T\right)+\log^4 T\ll\log^4 T.\end{aligned}$$

To estimate I_{31}, we use (2-9) to write

$$\begin{aligned}I_{31}=\frac{1}{2\pi i}\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT}x^{-s}\frac{\zeta'}{\zeta}(1-s)\,ds\\+\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT}x^{-s}\log\frac{t}{2\pi}\,ds+O\left(\int_{T_0}^{T}\frac{dt}{t}\right).\end{aligned}$$

Integrating by parts, we see that

$$\int_{T_0}^{T}x^{-it}\log\frac{t}{2\pi}\,dt\ll\frac{\log T}{\log x}.$$

Hence

$$I_{31}=\frac{1}{2\pi i}\int_{-1/\log(3x)+iT_0}^{-1/\log(3x)+iT}x^{-s}\frac{\zeta'}{\zeta}(1-s)\,ds+O\left(\frac{\log T}{\log x}\right)+O(\log T).$$

The remaining integral equals

$$\begin{aligned}&-\frac{1}{x}\sum_{n=2}^{\infty}\Lambda(n)\left(\frac{x}{n}\right)^{1+1/\log(3x)}\left(\frac{1}{2\pi}\int_{T_0}^{T}\left(\frac{x}{n}\right)^{-it}dt\right)\\&\quad=-\frac{T-T_0}{2\pi x}\Lambda(x)+O\left(\frac{1}{x}\sum_{\substack{n=2\\n\neq x}}^{\infty}\Lambda(n)\left(\frac{x}{n}\right)^{1+1/\log(3x)}\min\left\{T,\frac{1}{|\log(x/n)|}\right\}\right).\end{aligned}$$

By Lemma 2 of [Gonek 1993] this equals

$$-\frac{T-T_0}{2\pi x}\Lambda(x)+O(\log(2x)\log\log(3x))+O\left(\log(2x)\min\left\{\frac{T}{x},\frac{1}{\langle x\rangle}\right\}\right).$$

Hence,

$$I_{3,1} = -\frac{T}{2\pi x}\Lambda(x) + O(\log(2x)\log\log(3x)) + O\Big(\log(2x)\min\Big\{\frac{T}{x}, \frac{1}{\langle x\rangle}\Big\}\Big) + O\Big(\frac{\log T}{\log x}\Big) + O(\log T).$$

Combining our estimates for $I_{3,1}$ and $I_{3,2}$, we obtain

$$I_3 = -\frac{T}{2\pi x}\Lambda(x) + O(\log(2x)\log\log(3x)) + O\Big(\log(2x)\min\Big\{\frac{T}{x}, \frac{1}{\langle x\rangle}\Big\}\Big) + O\Big(\frac{\log T}{\log x}\Big) + O(\log^4 T). \tag{3-11}$$

Finally, since $\zeta'(s)/(\zeta(s)-a)$ is bounded on $[-2+iT_0,\ \sigma^*+1+iT_0]$,

$$I_4 \ll x^{1/\log(3x)} \ll 1.$$

It follows from this, (3-1), (3-2), and (3-11) that

$$\sum_{\substack{T_0<\gamma_a<T\\ \beta_a>0}} x^{-\rho_a} = -\frac{T}{2\pi x}\Lambda(x) + O\Big(\frac{\log T}{\log x}\Big) + O\Big(\log(2x)\min\Big\{\frac{T}{x}, \frac{1}{\langle x\rangle}\Big\}\Big) + O(\log^4 T). \tag{3-12}$$

To complete the proof of the theorem, we argue in much the same way as at the end of the proof of Theorem 1.1. That is, we first remove the constraint that no γ_a equals T and then note that we may extend the sum on the left-hand side of (3-12) to include the a-points with $0 < \gamma_a \le T_0$. Finally, it is easy to see that we may replace our condition that $T > T_0$ by $T > 1$.

4. Proof of Theorem 1.3

Levinson [1975] has shown that for $\delta > 0$ and T sufficiently large (depending on a), the number of a-points $\rho_a = \beta_a + i\gamma_a$ with $\big|\beta_a - \frac{1}{2}\big| > \delta$ and $T \le \gamma_a \le 2T$ is $O(\delta^{-1}T\log\log T)$. Thus,

$$\sum_{T<\gamma_a\le 2T}\big|\beta_a-\tfrac12\big| = \sum_{\substack{T<\gamma_a\le 2T\\ |\beta_a-1/2|>\delta}}\big|\beta_a-\tfrac12\big| + \sum_{\substack{T<\gamma_a\le 2T\\ |\beta_a-1/2|\le\delta}}\big|\beta_a-\tfrac12\big| \ll \frac{T\log\log T}{\delta} + \delta N_a(T).$$

Taking $\delta = (\log\log T/\log T)^{1/2}$, we deduce that

$$\sum_{T<\gamma_a\le 2T}\big|\beta_a-\tfrac12\big| \ll T\sqrt{\log T\log\log T}. \tag{4-1}$$

Since $e^y - 1 \ll |y| \max\{1, e^y\}$ for any $y > 0$, we see that

$$|x^{-1/2} - x^{-\beta_a}| = x^{-1/2}|1 - x^{1/2-\beta_a}| \ll \left|\beta_a - \tfrac{1}{2}\right| |\log x| \max\{x^{-1/2}, x^{-\beta_a}\}.$$

By the remark after (2-3), there is a number T_0 such that if $\gamma_a \geq T_0$, then $\beta_a > 0$. We may obviously also assume that T_0 is so large that (4-1) holds for $T \geq T_0$. It follows that if $x > 1$, then for these ρ_a we have $x^{-\beta_a} < 1$. Hence, for $x > 1$

$$|x^{-1/2} - x^{-\beta_a}| \ll \left|\beta_a - \tfrac{1}{2}\right| \log x.$$

This and (4-1) imply that

$$\sum_{T<\gamma_a\leq 2T} x^{-1/2-i\gamma_a} = \sum_{T<\gamma_a\leq 2T} x^{-\rho_a} + O\big(T \log x \sqrt{\log T \log\log T}\big)$$

for $x > 1$. Replacing T by $\frac{T}{2}, \frac{T}{4}, \frac{T}{8}, \ldots$ and summing, we see that

$$\sum_{T_0<\gamma_a\leq T} x^{-1/2-i\gamma_a} = \sum_{T_0<\gamma_a\leq T} x^{-\rho_a} + O\big(T \log x \sqrt{\log T \log\log T}\big).$$

Now fix $0 < \theta < 1$ and assume that $1 < x \leq T^\theta$. From this and Theorem 1.2 we find that

$$\begin{aligned}(4\text{-}2)\qquad \sum_{T_0<\gamma_a\leq T} x^{-i\gamma_a} &= -\frac{T}{2\pi\sqrt{x}}\Lambda(x) + O\Big(\sqrt{x}\frac{\log T}{\log x}\Big) + O(\sqrt{x}\log^4 T)\\ &\quad + O\Big(\sqrt{x}\log(2x)\min\Big\{\frac{T}{x}, \frac{1}{\langle x\rangle}\Big\}\Big) + O\big(\sqrt{x}T\log x\sqrt{\log T\log\log T}\big).\end{aligned}$$

By the Erdős–Turán inequality (see [Montgomery 1994, Chapter 1, Corollary 1.1]), if K is a positive integer, $\lambda \neq 0$ is a real number, and $[\alpha, \beta]$ is a subinterval of $[0, 1]$, then

$$\begin{aligned}(4\text{-}3)\qquad &\Bigg|\sum_{\substack{T_0<\gamma_a\leq T\\ \{\lambda\gamma_a\}\in[\alpha,\beta]}} 1 - (\beta-\alpha)(N_a(T) - N_a(T_0))\Bigg|\\ &\qquad\leq \frac{N_a(T)}{K+1} + 3\sum_{k\leq K}\frac{1}{k}\Bigg|\sum_{T_0<\gamma_a\leq T} e(k\lambda\gamma_a)\Bigg|.\end{aligned}$$

Without loss of generality we may assume that $\lambda > 0$. Taking $x = \exp(2\pi k\lambda)$ with k a positive integer in (4-2), and then taking the complex conjugates of both sides of the resulting equation, we find that

$$\frac{1}{k}\sum_{T_0<\gamma_a\leq T} e(k\lambda\gamma_a) \ll_\lambda \frac{T}{e^{\pi k\lambda}} + e^{\pi k\lambda}T\sqrt{\log T\log\log T}.$$

Inserting this into (4-3) and evaluating, we obtain

$$\Bigg|\sum_{\substack{T_0<\gamma_a\le T\\ \{\lambda\gamma_a\}\in[\alpha,\beta]}} 1-(\beta-\alpha)(N_a(T)-N_a(T_0))\Bigg| \ll \frac{N_a(T)}{K}+TK+\ e^{\pi K\lambda}T\sqrt{\log T\log\log T}.$$

Note that including the terms (if any) with $0<\gamma_a\le T_0,\ \beta_a>0$, and $\{\lambda\gamma_a\}\in[\alpha,\beta]$ changes the left-hand side by at most $O(1)$. If we now choose

$$K=\left[\frac{\frac{1}{2}-\varepsilon}{\pi\lambda}(\log\log T)\right],$$

we obtain

$$\left|\frac{1}{N_a(T)}\sum_{\substack{0<\gamma_a\le T,\ \beta_a>0\\ \{\lambda\gamma_a\}\in[\alpha,\beta]}} 1-(\beta-\alpha)\right| \ll \frac{1}{\log\log T}$$

for $\lambda>0$ fixed, and uniformly for any subinterval $[\alpha,\beta]$ of $[0,1]$. The estimate (1-6) follows easily from this.

5. Proof of Theorem 1.4 and Corollary 1.5

By (1-1) and Theorem 1.2 with $x=n$ an integer ≥ 2, we see that

$$\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}}\frac{1}{n^{\rho_a}}=-\frac{T}{2\pi n}\Lambda(n)+O(\log n)+O(\log^4 T).$$

Thus, since $N=T^\theta$ with $0<\theta<1$ fixed, we have

$$\begin{aligned}(5\text{-}1)\qquad \sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}}A(\rho_a)&=\sum_{n\le N}a(n)\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}}n^{-\rho_a}\\ &=a(1)N_a(T)+\sum_{2\le n\le N}a(n)\Big(-\frac{T}{2\pi n}\Lambda(n)+O(\log^4 T)\Big)\\ &=\frac{T}{2\pi}\left(a(1)\log T-\sum_{2\le n\le N}\frac{a(n)\Lambda(n)}{n}\right)+\ O(T^{\theta+2\varepsilon}).\end{aligned}$$

This gives Theorem 1.4, assuming ε is so small that $\theta+2\varepsilon\le 1$.

To prove Corollary 1.5, first take $A(s)=M(s)$ in (5-1), where

$$M(s)=\sum_{n\le N}\mu(n)n^{-s}$$

and $N = T^\theta$ with $0 < \theta < 1$ fixed. Then we find that

$$\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}} M(\rho_a) = \frac{T}{2\pi}\left(\log T - \sum_{2\le n\le N}\frac{\mu(n)\Lambda(n)}{n}\right) + O(T).$$

The sum over n equals

$$-\sum_{p\le N}\frac{\log p}{p} = -\log N + O(1).$$

Thus,

$$\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}} M(\rho_a) = \frac{T}{2\pi}\log T + \theta\frac{T}{2\pi}\log T + O(T),$$

which is the same as (1-7).

For $P(s) = \sum_{n\le N} n^{-s}$, we similarly find that

$$\begin{aligned}\sum_{\substack{0<\gamma_a\le T\\ \beta_a>0}} P(\rho_a) &= \frac{T}{2\pi}\left(\log T - \sum_{2\le n\le N}\frac{\Lambda(n)}{n}\right) + O(T)\\ &= \frac{T}{2\pi}(\log T - \log N) + O(T).\end{aligned}$$

This gives (1-8).

6. Proof of Theorems 1.6 and 1.7

As in the previous sections we assume $a \neq 0$ is a fixed complex number. Throughout this section we write

$$f(s) = \zeta(s) - a.$$

As we shall show, when $a \neq 1$ and σ is sufficiently large, $1/f(s)$ has a Dirichlet series representation

$$\frac{1}{f(s)} = \frac{1}{\zeta(s)-a} = \sum_{n=1}^{\infty}\frac{b_a(n)}{n^s}.$$

We shall also show that when $a = 1$ and σ is large, one has the generalized Dirichlet series representation

$$\frac{1}{f(s)} = \frac{1}{\zeta(s)-1} = \sum_{\substack{m=1\\ \text{odd}}}^{\infty}\sum_{r=-\infty}^{\infty}\frac{b_1(m2^r)}{(m2^r)^s}. \tag{6-1}$$

We denote the abscissa of convergence of $1/f(s)$ by σ_0 and its abscissa of absolute convergence by $\bar{\sigma}$. Both, of course, depend on a and, in general, neither is easy to determine precisely. Theorem 1.6 gives explicit formulae for the coefficients $b_a(n)$

of $f(s)$ and Theorem 1.7 gives upper bounds for $\bar{\sigma}$. The two theorems are most conveniently proved together for the various ranges of a.

First we consider the case when $|a| > 1$. Clearly $\zeta(\sigma)$ decreases from ∞ to 1 as σ increases from 1 to ∞. Hence $\zeta(\sigma) = |a|$ has a unique solution $\sigma^* > 1$, and for $\sigma > \sigma^*$ we have $\zeta(\sigma) < \zeta(\sigma^*)$. Moreover, $|\zeta(s)| \le \zeta(\sigma)$ for $\sigma > 1$. Thus, when $\sigma > \sigma^*$

$$|\zeta(s)| \le \zeta(\sigma) < \zeta(\sigma^*) = |a|.$$

Furthermore, for $\sigma > \sigma^*$ we have

$$\frac{1}{\zeta(s)-a} = -\frac{1}{a}\sum_{k=0}^{\infty}\left(\frac{\zeta(s)}{a}\right)^k = -\frac{1}{a}\sum_{k=0}^{\infty}\frac{1}{a^k}\sum_{n=1}^{\infty}\frac{d_k(n)}{n^s}.$$

The double sum, in fact, converges absolutely since $d_k(n)$ is positive and

$$-\frac{1}{|a|}\sum_{k=0}^{\infty}\frac{1}{|a|^k}\sum_{n=1}^{\infty}\frac{d_k(n)}{n^\sigma} = \frac{1}{\zeta(\sigma)-|a|}.$$

Thus, when $|a| > 1$ we have $\bar{\sigma} \le \sigma^*$ and

$$b_a(n) = -\sum_{k=0}^{\infty}\frac{d_k(n)}{a^{k+1}}.$$

Remark. It is not difficult to see from the proof that when $a > 1$ is real, we in fact have $\bar{\sigma} = \sigma^*$.

Next we consider the case $0 < |a| < 1$. For $\sigma > 1$

$$|\zeta(s)| \ge \prod_p\left(1+\frac{1}{p^\sigma}\right)^{-1} = \frac{\zeta(2\sigma)}{\zeta(\sigma)}.$$

Since $\zeta(2\sigma)/\zeta(\sigma)$ increases from 0 to 1 as σ increases from 1 to ∞, there is a unique solution $\sigma^* > 1$ of the equation $\zeta(2\sigma)/\zeta(\sigma) = |a|$, and if $\sigma > \sigma^*$, then $|\zeta(s)| \ge \zeta(2\sigma)/\zeta(\sigma) > |a|$. Thus, for $\sigma > \sigma^*$

$$\frac{1}{\zeta(s)-a} = \sum_{k=0}^{\infty}\frac{a^k}{\zeta(s)^{k+1}} = \sum_{k=0}^{\infty}a^k\sum_{n=1}^{\infty}\frac{d_{-(k+1)}(n)}{n^s}. \tag{6-2}$$

For any prime power p^j, we have $d_{-(k+1)}(p^j) = \binom{k+1}{j}(-1)^j$. Hence

$$\sum_{n=1}^{\infty}\frac{|d_{-(k+1)}(n)|}{n^\sigma} = \prod_p\left(\sum_{j=0}^{k+1}\binom{k+1}{j}\frac{1}{p^{j\sigma}}\right) = \left(\frac{\zeta(\sigma)}{\zeta(2\sigma)}\right)^{k+1}.$$

Therefore the double sum in (6-2) is absolutely convergent and has modulus

$$\leq \sum_{k=0}^{\infty} |a|^k \left(\frac{\zeta(\sigma)}{\zeta(2\sigma)} \right)^{k+1} = \frac{\zeta(\sigma)}{\zeta(2\sigma)} \cdot \frac{1}{1 - |a|\zeta(\sigma)/\zeta(2\sigma)}.$$

It follows that $\bar{\sigma} \leq \sigma^*$ and that

$$b_a(n) = \sum_{k=1}^{\infty} a^{k-1} d_{-k}(n).$$

Suppose next that $|a| = 1$ but $a \neq 1$. If $\sigma > 1$

$$|\zeta(s) - 1| \leq \zeta(\sigma) - 1,$$

and the right-hand side decreases from ∞ to 0 as σ increases from 1 to ∞. Thus, there is a unique solution $\sigma^* > 1$ to the equation $\zeta(\sigma) - 1 = |a - 1|$. Moreover, if $\sigma > \sigma^*$, then $|\zeta(s) - 1| < \zeta(\sigma^*) - 1 = |a - 1|$. Hence, for $\sigma > \sigma^*$,

$$|\zeta(s) - 1| < |a - 1|.$$

We therefore see that

$$\frac{1}{\zeta(s) - a} = \frac{1}{(\zeta(s) - 1) - (a - 1)} = -\sum_{k=0}^{\infty} \frac{(\zeta(s) - 1)^k}{(a - 1)^{k+1}}$$

$$= -\sum_{k=0}^{\infty} \frac{1}{(a - 1)^{k+1}} \sum_{n=1}^{\infty} \frac{e_k(n)}{n^s},$$

where

$$(\zeta(s) - 1)^k = \sum_{n=1}^{\infty} \frac{e_k(n)}{n^s} \quad (\sigma > 1). \tag{6-3}$$

We note that the $e_k(n) \geq 0$, so for $\sigma > \sigma^*$

$$\sum_{k=0}^{\infty} \frac{1}{|a - 1|^{k+1}} \sum_{n=1}^{\infty} \frac{e_k(n)}{n^{\sigma}} = \sum_{k=0}^{\infty} \frac{(\zeta(\sigma) - 1)^k}{|a - 1|^{k+1}} = \frac{1}{|a - 1| - (\zeta(\sigma) - 1)}.$$

Thus $\bar{\sigma} \leq \sigma^*$, where σ^* is the unique solution to $\zeta(\sigma) = 1 + |1 - a|$ in $\sigma > 1$. We also see that

$$b_a(n) = -\sum_{k=0}^{\infty} \frac{e_k(n)}{(a - 1)^{k+1}},$$

where $e_k(n)$ is given by (6-3).

Finally, suppose that $a = 1$. Then for $\sigma > 1$

$$\frac{1}{\zeta(s) - 1} = \frac{2^s}{1 + \left(\frac{2}{3}\right)^s + \left(\frac{2}{4}\right)^s + \cdots}.$$

This time we let σ^* be the unique solution in $\sigma > 1$ of

$$1 = \left(\tfrac{2}{3}\right)^\sigma + \left(\tfrac{2}{4}\right)^\sigma + \cdots$$

or, equivalently, of

$$\zeta(\sigma) = 1 + 2^{1-\sigma}.$$

Then if $\sigma > \sigma^*$,

$$\left(\tfrac{2}{3}\right)^\sigma + \left(\tfrac{2}{4}\right)^\sigma + \cdots < 1$$

and we have

$$\begin{aligned}\frac{1}{\zeta(s)-1} &= \frac{2^s}{1+\left(\frac{2}{3}\right)^s + \left(\frac{2}{4}\right)^s + \cdots} \\ &= \sum_{k=0}^{\infty} (-1)^k 2^{(k+1)s} \left(\zeta(s) - 1 - \frac{1}{2^s}\right)^k \\ &= \sum_{k=0}^{\infty} (-1)^k 2^{(k+1)s} \sum_{n=3^k}^{\infty} \frac{f_k(n)}{n^s},\end{aligned} \tag{6-4}$$

where

$$(\zeta(s) - 1 - 2^{-s})^k = \sum_{n=3^k}^{\infty} \frac{f_k(n)}{n^s}.$$

By our choice of σ^*, the double series in (6-4) converges absolutely when $\sigma > \sigma^*$. Thus we have $\bar{\sigma} \le \sigma^*$, and (6-1) holds with coefficients given by

$$b_1(m2^r) = \sum_{\substack{l-k-1=r \\ k,l \ge 0}} (-1)^k f_k(2^l m).$$

This completes our proof of Theorems 1.6 and 1.7.

7. Proof of Theorem 1.8

By a theorem of Landau [1933, Appendix, Satz 12], if

$$g(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

is convergent and nonzero for $\sigma > \alpha$ and $a_1 \neq 0$, then

$$\frac{1}{g(s)} = \sum_{n=1}^{\infty} \frac{b_n}{n^s}$$

converges for $\sigma > \alpha$. We apply this to the function $g(s) = \zeta(s) - a$ when $a \neq 1$ or 0. Let $\rho_a = \beta_a + i\gamma_a$ denote a typical zero of $\zeta(s) - a$ and let

$$\beta_a^* = \sup_{\rho_a} \beta_a,$$

as before. Then the series for $1/(\zeta(s) - a)$ converges when $\sigma > \beta_a^*$. In fact, β_a^* is the exact abscissa of convergence because $1/(\zeta(s) - a)$ has a pole at every zero ρ_a of $\zeta(s) - a$ and, therefore, the series cannot converge at ρ_a. Thus, we have $\sigma_0 = \beta_a^* \leq \bar{\sigma}$. Next recall that $\beta_a^* > 1$ (see just after (1-5)). From this and Theorem 1.7 we see that for $a \neq 0, 1$,

$$1 < \sigma_0 = \beta_a^* \leq \bar{\sigma} \leq \sigma^*.$$

Finally we turn to the growth of the coefficients $b_a(n)$. Since the terms $|b_a(n)/n^\sigma|$ must tend to zero when $\sigma > \beta_a^*$, it is clear that for any $\varepsilon > 0$

$$b_a(n) \ll n^{\beta_a^* + \varepsilon}.$$

By a theorem of Bombieri and Ghosh [2011, Theorem 3] this upper bound is sharp when $a \neq 0, 1$ in the sense that

$$|b_a(n)| > n^{\beta_a^* - \varepsilon}$$

for infinitely many n. This completes the proof of Theorem 1.8.

Acknowledgment

We thank the referee for suggesting that it might be possible to strengthen one of our error estimates in the proof of Theorem 1.2. The improvement allowed us to increase the range of θ in Theorem 1.4 and Corollary 1.5. We are also very grateful to Athanasios Sourmelidis for noticing an error in an earlier version of this paper and suggesting a way to correct it.

References

[Bohr and Landau 1914] H. Bohr and E. Landau, "Ein Satz über Dirichletsche Reihen mit Andwendung auf die ζ-Funktion und die L-Funktionen", *Palermo Rend.* **37**:1 (1914), 269–272. Zbl

[Bohr et al. 1913] H. Bohr, E. Landau, and J. E. Littlewood, "Sur la fonction $\zeta(s)$ dans le voisinage de la droite $\sigma = 1/2$", *Belg. Bull. Sc.* **15** (1913), 1144–1175. Zbl

[Bombieri and Ghosh 2011] E. Bombieri and A. Ghosh, "On the Davenport–Heilbronn function", *Uspekhi Mat. Nauk* **66**:2(398) (2011), 15–66. In Russian; translated in *Russ. Math. Surv.* **66**:2 (2011), 221–270. Erratum in *Uspekhi Mat. Nauk* **66**:3(399) (2011), 208. MR Zbl

[Davenport 1980] H. Davenport, *Multiplicative number theory*, 2nd ed., Graduate Texts in Math. **74**, Springer, 1980. MR Zbl

[Gonek 1984] S. M. Gonek, "Mean values of the Riemann zeta function and its derivatives", *Invent. Math.* **75**:1 (1984), 123–141. MR Zbl

[Gonek 1985] S. M. Gonek, "A formula of Landau and mean values of $\zeta(s)$", pp. 92–97 in *Topics in analytic number theory* (Austin, TX, 1982), edited by S. W. Graham and J. D. Vaaler, Univ. Texas Press, Austin, TX, 1985. MR Zbl

[Gonek 1993] S. M. Gonek, "An explicit formula of Landau and its applications to the theory of the zeta-function", pp. 395–413 in *A tribute to Emil Grosswald: number theory and related analysis*, edited by M. Knopp and M. Sheingorn, Contemp. Math. **143**, Amer. Math. Soc., Providence, RI, 1993. MR Zbl

[Landau 1912] E. Landau, "Über die Nullstellen der Zetafunktion", *Math. Ann.* **71**:4 (1912), 548–564. MR Zbl

[Landau 1933] E. Landau, "Über den Wertevorrat von $\zeta(s)$ in der Halbebene $\sigma > 1$", *Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl.* **36** (1933), 81–91. Zbl

[Levinson 1975] N. Levinson, "Almost all roots of $\zeta(s) = a$ are arbitrarily close to $\sigma = 1/2$", *Proc. Nat. Acad. Sci. U.S.A.* **72** (1975), 1322–1324. MR Zbl

[Montgomery 1994] H. L. Montgomery, *Ten lectures on the interface between analytic number theory and harmonic analysis*, CBMS Region. Conf. Series in Math. **84**, Amer. Math. Soc., Providence, RI, 1994. MR Zbl

[Selberg 1992] A. Selberg, "Old and new conjectures and results about a class of Dirichlet series", pp. 367–385 in *Proceedings of the Amalfi Conference on Analytic Number Theory* (Maiori, Italy, 1989), edited by E. Bombieri et al., Univ. Salerno, 1992. MR Zbl

[Steuding 2014] J. Steuding, "One hundred years uniform distribution modulo one and recent applications to Riemann's zeta-function", pp. 659–698 in *Topics in mathematical analysis and applications*, edited by T. M. Rassias and L. Tóth, Springer Optim. Appl. **94**, Springer, 2014. MR Zbl

[Titchmarsh 1932] E. C. Titchmarsh, *The theory of functions*, Clarendon Press, Oxford, 1932. Zbl

[Titchmarsh 1986] E. C. Titchmarsh, *The theory of the Riemann zeta-function*, 2nd ed., Oxford Univ. Press, 1986. MR Zbl

Received January 7, 2019. Revised May 5, 2019.

SIEGFRED BALUYOT
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ROCHESTER
ROCHESTER, NY
UNITED STATES
Current address:
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
URBANA, IL
UNITED STATES
sbaluyot@illinois.edu

STEVEN M. GONEK
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ROCHESTER
ROCHESTER, NY
UNITED STATES
gonek@math.rochester.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.73

DIFFEOLOGICAL VECTOR SPACES

J. DANIEL CHRISTENSEN AND ENXIN WU

We study the relationship between many natural conditions that one can put on a diffeological vector space: being fine or projective, having enough smooth (or smooth linear) functionals to separate points, having a diffeology determined by the smooth linear functionals, having fine finite-dimensional subspaces, and having a Hausdorff underlying topology. Our main result is that the majority of the conditions fit into a total order. We also give many examples in order to show which implications do not hold, and use our results to study the homological algebra of diffeological vector spaces.

1. Introduction

Diffeological spaces are elegant generalizations of manifolds that include a variety of singular spaces and infinite-dimensional spaces. Many vector spaces that arise in applications are naturally equipped with a compatible structure of a diffeological space. Examples include $C^\infty(M, \mathbb{R}^n)$ for a manifold (or even a diffeological space) M, spaces of smooth or holomorphic sections of vector bundles, tangent spaces of diffeological spaces (as defined in [Christensen and Wu 2016]), smooth linear duals of all of these spaces, etc. Such objects are called diffeological vector spaces and are the topic of this paper.

Diffeological vector spaces have been studied by Iglesias-Zemmour [2007; 2013]. He used them to define diffeological manifolds, and developed the theory of *fine* diffeological vector spaces, a particularly well-behaved kind that forms the beginning of our story. Kriegl and Michor [1997] studied topological vector spaces equipped with a smooth structure, and their examples can be regarded as diffeological vector spaces. Diffeological vector spaces were used in the study of tangent spaces of diffeological spaces in [Vincent 2008] and [Christensen and Wu 2016]. Wu [2015] investigated the homological algebra of *all* diffeological vector spaces and the present paper builds heavily on this foundation.

The second author was partially supported by NNSF of China (No. 112530) and STU Scientific Research Foundation for Talents (No. 760179).

MSC2010: primary 46S99; secondary 57P99.

Keywords: diffeological vector space, homological algebra, fine diffeology, projective diffeological vector space, smooth linear functionals.

In this paper, we study some natural conditions that one can put on a diffeological vector space, and show that the majority of them fit into a total order. In order to state our results, we briefly introduce the conditions here, making use of some background material summarized in Section 2.

Any vector space has a smallest diffeology making it into a diffeological vector space. This is called the *fine* diffeology, and we write $\mathcal{FV}$ for the collection of vector spaces with the fine diffeology. We write $\mathcal{FFV}$ for the collection of diffeological vector spaces whose finite-dimensional subspaces (with the induced diffeology) are all fine.

A diffeological vector space V is *projective* if for every linear subduction $f : W_1 \to W_2$ and every smooth linear map $g : V \to W_2$, there exists a smooth linear map $h : V \to W_1$ such that $g = f \circ h$. We write $\mathcal{PV}$ for the collection of projective diffeological vector spaces.

A diffeological vector space V is in $\mathcal{SD}$ (resp. $\mathcal{SV}$) if the smooth (resp. smooth linear) functionals $V \to \mathbb{R}$ separate points of V. That is, for each x and y in V with $x \neq y$, such a functional f can be found so that $f(x) \neq f(y)$.

Each diffeological space has a natural topology called the D-topology. We write $\mathcal{HT}$ for the collection of diffeological vector spaces whose D-topologies are Hausdorff.

The last letter of the abbreviation is $\mathcal{V}$, $\mathcal{D}$ or $\mathcal{T}$ depending on whether the condition depends on the structure as a diffeological vector space, a diffeological space, or a topological space.

We now state the main results of the paper.

Theorem 1.1. *We have the following chain of containments*:

$$\mathcal{FV} \subset \mathcal{PV} \subset \mathcal{SV} \subseteq \mathcal{SD} \subset \mathcal{FFV} \quad \textit{and} \quad \mathcal{SD} \subset \mathcal{HT},$$

where $\subset$ *indicates proper containment. Neither of* $\mathcal{HT}$ *and* $\mathcal{FFV}$ *contains the other.*

We do not know whether the containments $\mathcal{SV} \subseteq \mathcal{SD}$ and $\mathcal{SD} \subseteq \mathcal{FFV} \cap \mathcal{HT}$ are proper.

The property of being finite-dimensional does not imply, nor is it implied by, any of the properties considered above. However, under this assumption, most of the properties agree.

Theorem 1.2. *When restricted to finite-dimensional vector spaces, the collections* $\mathcal{FV}$, $\mathcal{PV}$, $\mathcal{SV}$, $\mathcal{SD}$ *and* $\mathcal{FFV}$ *agree.*

Indeed, $\mathcal{FV}$ and $\mathcal{FFV}$ clearly agree for finite-dimensional spaces, so the containments must collapse to equalities. Note that we prove part of Theorem 1.2 (see Theorem 3.19) on the way to proving Theorem 1.1.

The final property we consider is the following. Write $\mathcal{DV}$ for the collection of diffeological vector spaces V such that a function $p : \mathbb{R}^n \to V$ is smooth if and only

if $\ell \circ p : \mathbb{R}^n \to \mathbb{R}$ is smooth for each smooth linear functional $\ell : V \to \mathbb{R}$. Except for the inclusion $\mathcal{FV} \subset \mathcal{DV}$, the class $\mathcal{DV}$ is independent of all of the others we have considered. However, under this assumption, we again find that many of the other conditions agree.

Theorem 1.3. *When restricted to V in $\mathcal{DV}$, the collections $\mathcal{SV}$, $\mathcal{SD}$, $\mathcal{FFV}$ and $\mathcal{HT}$ agree.*

The proofs of the containments, and the examples showing that many inclusions do not hold, are spread throughout Section 3. For example, we show $\mathcal{FV} \subset \mathcal{PV}$ in Example 3.7 and Proposition 3.8, $\mathcal{PV} \subset \mathcal{SV}$ in Proposition 3.14 and Remark 3.15 (1), $\mathcal{SD} \subset \mathcal{HT}$ in Proposition 3.16 and Example 3.18, $\mathcal{HT} \nsubseteq \mathcal{FFV}$ in Example 3.18, and both $\mathcal{SD} \neq \mathcal{FFV}$ and $\mathcal{FFV} \nsubseteq \mathcal{HT}$ in Proposition 3.23. That $\mathcal{PV} \nsubseteq \mathcal{DV}$ is Proposition 3.33, and the proof of Theorem 1.3 is in Proposition 3.31. The longest argument, which is the proof that $\mathcal{SD} \subseteq \mathcal{FFV}$, is deferred until Section 5. Along the way, we also prove other results, such as the fact that a diffeological vector space V is fine if and only if every linear functional on V is smooth, and some necessary conditions for diffeological vector spaces and free diffeological vector spaces to be projective. In Section 4, we give some applications of our results to the homological algebra of diffeological vector spaces. For example, we show that every finite-dimensional subspace of a diffeological vector space in $\mathcal{SV}$ is a smooth direct summand.

We are thankful to Chengjie Yu for the argument used in Case 1 of the proof of Theorem 3.22 in Section 5 and to the referee for many comments that helped improve the exposition.

2. Background and conventions

In this section, we briefly recall some background on diffeological spaces. For further details, we recommend the standard textbook [Iglesias-Zemmour 2013]. For a concise introduction to diffeological spaces, we recommend [Christensen et al. 2014], particularly Section 2 and the introduction to Section 3.

Definition 2.1 [Souriau 1984]. A *diffeological space* is a set X together with a specified set of functions $U \to X$ (called *plots*) for each open set U in $\mathbb{R}^n$ and each $n \in \mathbb{N}$, such that for all open subsets $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$:

(1) (covering) Every constant function $U \to X$ is a plot.

(2) (smooth compatibility) If $U \to X$ is a plot and $V \to U$ is smooth, then the composite $V \to U \to X$ is also a plot.

(3) (sheaf condition) If $U = \bigcup_i U_i$ is an open cover and $U \to X$ is a function such that each restriction $U_i \to X$ is a plot, then $U \to X$ is a plot.

A function $f : X \to Y$ between diffeological spaces is *smooth* if for every plot $p : U \to X$ of X, the composite $f \circ p$ is a plot of Y.

The category of diffeological spaces and smooth maps is complete and cocomplete. Given two diffeological spaces X and Y, we write $C^\infty(X, Y)$ for the set of all smooth maps from X to Y. An isomorphism in the category of diffeological spaces will be called a *diffeomorphism*.

Every manifold M is canonically a diffeological space with the plots taken to be all smooth maps $U \to M$ in the usual sense. We call this the *standard diffeology* on M. It is easy to see that smooth maps in the usual sense between manifolds coincide with smooth maps between them with the standard diffeology.

For a diffeological space X with an equivalence relation $\sim$, the *quotient diffeology* on $X/\sim$ consists of all functions $U \to X/\sim$ that locally factor through the quotient map $X \to X/\sim$ via plots of X. A *subduction* is a map diffeomorphic to a quotient map. That is, it is a map $X \to Y$ such that the plots in Y are the functions that locally lift to X as plots in X.

For a diffeological space Y and a subset A of Y, the *subdiffeology* consists of all functions $U \to A$ such that $U \to A \hookrightarrow Y$ is a plot of Y. An *induction* is an injective smooth map $A \to Y$ such that a function $U \to A$ is a plot of A if and only if $U \to A \to Y$ is a plot of Y.

For diffeological spaces X and Y, the *product diffeology* on $X \times Y$ consists of all functions $U \to X \times Y$ whose components $U \to X$ and $U \to Y$ are plots of X and Y, respectively.

The *discrete diffeology* on a set is the diffeology whose plots are the locally constant functions. The *indiscrete diffeology* on a set is the diffeology in which every function is a plot.

We can associate to every diffeological space the following topology:

Definition 2.2 [Iglesias-Zemmour 2007]. Let X be a diffeological space. A subset A of X is *D-open* if $p^{-1}(A)$ is open in U for each plot $p : U \to X$. The collection of D-open subsets of X forms a topology on X called the *D-topology*.

Definition 2.3. A *diffeological vector space* is a vector space V with a diffeology such that addition $V \times V \to V$ and scalar multiplication $\mathbb{R} \times V \to V$ are smooth.

Let V be a diffeological vector space. We write $L^\infty(V, \mathbb{R})$ for the set of all smooth linear maps $V \to \mathbb{R}$, and $L(V, \mathbb{R})$ for the set of all linear maps $V \to \mathbb{R}$. We write DVect for the category of diffeological vector spaces and smooth linear maps.

Conventions. Throughout this paper, we use the following conventions. Every subset of a diffeological space is equipped with the subdiffeology and every product is equipped with the product diffeology. Every vector space is over the field $\mathbb{R}$ of real numbers, and every linear map is $\mathbb{R}$-linear. By a subspace of a diffeological

vector space, we mean a linear subspace with the subdiffeology. All manifolds are smooth, finite-dimensional, Hausdorff, second countable and without boundary, and are equipped with the standard diffeology.

3. Diffeological vector spaces

In this section, we study a variety of conditions that a diffeological vector space can satisfy. Together, the results described here give the theorems stated in the introduction. In addition, we present some auxiliary results, and give many examples and counterexamples.

Fine diffeological vector spaces. In this subsection, we recall background on the fine diffeology, and then give two new characterizations.

Given a vector space V, the set of all diffeologies on V each of which makes V into a diffeological vector space, ordered by inclusion, is a complete lattice. This follows from [Christensen and Wu 2016, Proposition 4.6], taking X to be a point. The largest element in this lattice is the indiscrete diffeology, which is usually not interesting. Another extreme has the following special name in the literature:

Definition 3.1. The *fine* diffeology on a vector space V is the smallest diffeology on V making it into a diffeological vector space.

For example, the fine diffeology on $\mathbb{R}^n$ is the standard diffeology.

Remark 3.2. The fine diffeology is generated by the injective linear maps $\mathbb{R}^n \to V$; see [Iglesias-Zemmour 2013, 3.8]. That is, the plots of the fine diffeology are the functions $p : U \to V$ such that for each $u \in U$, there are an open neighbourhood W of u in U, an injective linear map $i : \mathbb{R}^n \to V$ for some $n \in \mathbb{N}$, and a smooth map $f : W \to \mathbb{R}^n$ such that $p|_W = i \circ f$.

One can show that if V is any diffeological vector space and $p : W \to V$ is a plot that factors smoothly through some linear injection $\mathbb{R}^n \to V$, then every factorization of p through a linear injection $\mathbb{R}^m \to V$ is smooth. It follows that every subspace of a fine diffeological vector space is fine; see [Wu 2015].

In fact, fineness of a diffeological vector space can be tested by smooth curves:

Proposition 3.3. *A diffeological vector space V is fine if and only if for every plot $p : \mathbb{R} \to V$ and every $x \in \mathbb{R}$, there exist an open neighbourhood W of x in $\mathbb{R}$, an injective linear map $i : \mathbb{R}^n \to V$ for some $n \in \mathbb{N}$, and a smooth map $f : W \to \mathbb{R}^n$ such that $p|_W = i \circ f$.*

Proof. ($\Rightarrow$) This follows from the description of the fine diffeology in Remark 3.2.

($\Leftarrow$) Under the given assumptions, we will prove that V is fine. Let $q : U \to V$ be a plot and let u be a point in U. We first show that there is an open neighbourhood W of u in U such that $q|_W$ lands in a finite-dimensional subspace of V. If not, then

there exists a sequence u_i in U converging to u such that $\{q(u_i) \mid i \in \mathbb{Z}^+\}$ is linearly independent in V. We may assume that the sequence u_i converges fast to u; see [Kriegl and Michor 1997, I.2.8]. By the special curve lemma [Kriegl and Michor 1997, I.2.8], there exists a smooth map $f : \mathbb{R} \to U$ such that $f(1/i) = u_i$ and $f(0) = u$. Then $q \circ f : \mathbb{R} \to V$ is a plot which does not satisfy the hypothesis at $x = 0$.

So let W be an open neighbourhood of u in U such that $q|_W$ factors as $i \circ g$, where $i : \mathbb{R}^m \to V$ is a linear injection and $g : W \to \mathbb{R}^m$ is a function. We will prove that g is smooth. By Boman's theorem (see, e.g., [Kriegl and Michor 1997, Corollary 3.14]), it is enough to show that $g \circ r$ is smooth for every smooth curve $r : \mathbb{R} \to W$. Since $i \circ g \circ r$ is smooth, our assumption implies that it locally factors smoothly through an injective linear map $\mathbb{R}^n \to V$. Then the last part of Remark 3.2 implies that $g \circ r$ is locally smooth, and therefore smooth, as required. □

Proposition 3.4. *A diffeological vector space V is fine if and only if $L^\infty(V, \mathbb{R}) = L(V, \mathbb{R})$, i.e., if and only if every linear functional is smooth.*

Proof. This follows from the proof of [Wu 2015, Proposition 5.7]. We give a direct proof here.

It is easy to check that if V is fine, then every linear functional is smooth.

To prove the converse, suppose that every linear functional $V \to \mathbb{R}$ is smooth. Let $p : U \to V$ be a plot and let $u \in U$. First we must show that when restricted to a neighbourhood of u, p lands in a finite-dimensional subspace of V. If not, then there is a sequence $\{u_j\}$ converging to u such that the vectors $p(u_j)$ are linearly independent. Thus there is a linear functional $l : V \to \mathbb{R}$ such that $p(u_j)$ is sent to 1 when j is odd and 0 when j is even. By assumption, l is smooth. But $l \circ p$ is not continuous, contradicting the fact that p is a plot.

So now we know that p locally factors through an injective linear map $i : \mathbb{R}^n \to V$. (Of course, n may depend on the neighbourhood.) For each $1 \le j \le n$, there is a linear map $l_j : V \to \mathbb{R}$ such that $l_j \circ i$ is projection onto the j-th coordinate. Since $l_j \circ p$ is smooth, it follows that the local factorizations through $\mathbb{R}^n$ are smooth. Thus V is fine. □

Projective diffeological vector spaces.

Definition 3.5. A diffeological vector space V is *projective* if for every linear subduction $f : W_1 \to W_2$ and every smooth linear map $g : V \to W_2$, there exists a smooth linear map $h : V \to W_1$ making the diagram

$$\begin{array}{ccc} & & W_1 \\ & {\scriptstyle h}\nearrow & \big\downarrow {\scriptstyle f} \\ V & \xrightarrow[g]{} & W_2 \end{array}$$

commute. We write $\mathcal{PV}$ for the collection of projective diffeological vector spaces.

We now describe what will be a recurring example in this paper.

Definition 3.6. The *free diffeological vector space generated by a diffeological space* X is the vector space $F(X)$ with basis consisting of the elements of X and with the smallest diffeology making it into a diffeological vector space and such that the natural map $X \to F(X)$ is smooth.

This has the universal property that for any diffeological vector space V, every smooth map $X \to V$ extends uniquely to a smooth linear map $F(X) \to V$. Also, every plot in $F(X)$ is locally of the form

$$u \mapsto \sum_{i=1}^{k} r_i(u)[p_i(u)]$$

for smooth functions $r_i : U \to \mathbb{R}$ and $p_i : U \to X$, where for $x \in X$, $[x]$ denotes the corresponding basis vector in $F(X)$. See [Wu 2015, Proposition 3.5] for details.

Example 3.7. By [Wu 2015, Corollary 6.4], when M is a manifold, $F(M)$ is projective. However, by [Wu 2015, Theorem 5.3], $F(X)$ is fine if and only if X is discrete. So not every projective diffeological vector space is fine.

Proposition 3.8 [Wu 2015, Corollary 6.3]. *Every fine diffeological vector space is projective.*

Proof. This follows immediately from Proposition 3.4. One can take h to be $k \circ g$, where k is a linear section of f (which is not necessarily smooth). □

Projective diffeological vector spaces and the homological algebra of diffeological vector spaces are studied further in [Wu 2015].

Separation of points.

Definition 3.9. Let X be a diffeological space. A set A of functions with domain X is said to *separate points* if for any $x, y \in X$ with $x \neq y$, there exists $f \in A$ such that $f(x) \neq f(y)$. We say that the *smooth functionals separate points* if $C^\infty(X, \mathbb{R})$ separates points. We write $\mathcal{SD}'$ for the collection of all such diffeological spaces X and $\mathcal{SD}$ for the diffeological vector spaces whose underlying diffeological spaces are in $\mathcal{SD}'$. If V is a diffeological vector space, we say that the *smooth linear functionals separate points* if $L^\infty(V, \mathbb{R})$ separates points, and we write $\mathcal{SV}$ for the collection of all such diffeological vector spaces V.

We establish basic properties of such diffeological vector spaces below, and show that many familiar diffeological vector spaces have this property. Clearly, $\mathcal{SV} \subseteq \mathcal{SD}$.

Example 3.10. Every fine diffeological vector space is in $\mathcal{SV}$, since the coordinate functions with respect to any basis are smooth and linear. Every manifold is in $\mathcal{SD}'$, since the products of local coordinates with bump functions separate points (or by Whitney's embedding theorem).

Proposition 3.11. (1) *If $W \to V$ is a smooth linear injective map between diffeological vector spaces and $V \in \mathcal{SV}$, then $W \in \mathcal{SV}$. In particular, $\mathcal{SV}$ is closed under taking subspaces.*

(2) *Let $\{V_i\}_{i\in I}$ be a set of diffeological vector spaces. Then $\prod_{i\in I} V_i \in \mathcal{SV}$ if and only if each $V_i \in \mathcal{SV}$.*

(3) *Let $\{V_i\}_{i\in I}$ be a set of diffeological vector spaces. Then $\bigoplus_{i\in I} V_i \in \mathcal{SV}$ if and only if each $V_i \in \mathcal{SV}$, where $\bigoplus_{i\in I} V_i$ is the coproduct in* DVect; *see* [*Wu 2015, Proposition* 3.2].

Proof. This is straightforward. □

Proposition 3.12. *If $V \in \mathcal{SV}$ and X is a diffeological space, then $C^\infty(X, V) \in \mathcal{SV}$.*

Proof. This follows from the fact that every evaluation map $C^\infty(X, V) \to V$ is smooth and linear. □

Proposition 3.13. *The following are equivalent*:

(1) $X \in \mathcal{SD}'$.

(2) $F(X) \in \mathcal{SV}$.

(3) $F(X) \in \mathcal{SD}$.

Proof. It is enough to prove (1) ⇒ (2), since (2) ⇒ (3) ⇒ (1) are straightforward. Let $v \in F(X)$ be nonzero. It suffices to show that there is a smooth linear functional $F(X) \to \mathbb{R}$ which is nonzero on v. Write $v = \sum_{i=1}^k r_i[x_i]$ with $k \geq 1$, r_i nonzero for each i, and the x_i distinct. Since $C^\infty(X, \mathbb{R})$ separates points of X, there exists $f \in C^\infty(X, \mathbb{R})$ such that $f(x_1) = 1$ and $f(x_i) = 0$ for each $i > 1$. By the universal property of $F(X)$, f extends to a smooth linear map $F(X) \to \mathbb{R}$ which sends v to r_1, which is nonzero. □

Proposition 3.14. *Every projective diffeological vector space is in $\mathcal{SV}$.*

Proof. By Example 3.10, every open subset U of a Euclidean space is in $\mathcal{SD}'$. So Proposition 3.13 implies that $F(U)$ is in $\mathcal{SV}$. Corollary 6.15 of [Wu 2015] says that every projective diffeological vector space is a retract of a coproduct of $F(U)$'s in DVect. Therefore, it follows from Proposition 3.11 (3) and (1) that every projective diffeological vector space is in $\mathcal{SV}$. □

Remark 3.15. (1) Not every diffeological vector space in $\mathcal{SV}$ is projective. For example, let $V := \prod_\omega \mathbb{R}$ be the product of countably many copies of $\mathbb{R}$. By Proposition 3.11 (2), V is in $\mathcal{SV}$. But [Wu 2015, Example 4.3] shows that V is not projective.

(2) $\mathcal{SV}$ is not closed under taking quotients in DVect. For example, $F(\pi) : F(\mathbb{R}) \to F(T_\alpha)$ is a linear subduction, where α is an irrational and

$$\pi : \mathbb{R} \to T_\alpha := \mathbb{R}/(\mathbb{Z} + \alpha\mathbb{Z})$$

is the projection to the 1-dimensional irrational torus. By Proposition 3.13, $F(\mathbb{R})$ is in $\mathcal{SV}$, but $F(T_\alpha)$ is not in $\mathcal{SV}$ since T_α is not in $\mathcal{SD}'$. In particular, the free diffeological vector space $F(T_\alpha)$ is not projective, as observed in [Wu 2015, Example 4.3].

Here is an easy fact:

Proposition 3.16. *The D-topology of every diffeological space in $\mathcal{SD}'$ is Hausdorff. In particular, $\mathcal{SD} \subseteq \mathcal{HT}$.*

Proof. This follows from the fact that every smooth map is continuous when both domain and codomain are equipped with the D-topology. □

Corollary 3.17. *If $F(X)$ is projective, then X is Hausdorff.*

Proof. If $F(X)$ is projective, then it is in $\mathcal{SD}$ by Proposition 3.14, and so X is in $\mathcal{SD}'$ by Proposition 3.13. Thus X is Hausdorff, by Proposition 3.16. □

This gives another proof that free diffeological vector spaces are not always projective. For example, if a set X with more than one point is equipped with the indiscrete diffeology, then the D-topology on X is indiscrete as well, and hence $F(X)$ is not projective.

Example 3.18. The converse of Proposition 3.16 does not hold. Write $C(\mathbb{R})$ for the vector space $\mathbb{R}$ equipped with the continuous diffeology, so that a function $p : U \to C(\mathbb{R})$ is a plot if and only if it is continuous; see [Christensen et al. 2014, Section 3]. Then $C(\mathbb{R})$ is a Hausdorff diffeological vector space, as the D-topology on $C(\mathbb{R})$ is the usual topology. But one can show that $C^\infty(C(\mathbb{R}), \mathbb{R})$ consists of constant functions ([Christensen and Wu 2016, Example 3.15]), so $C(\mathbb{R})$ is not in $\mathcal{SD}$.

We will use the following result in the next subsection.

Theorem 3.19. *Let V be a finite-dimensional diffeological vector space. Then the following are equivalent*:

(1) *V is fine.*

(2) *V is projective.*

(3) *V is in $\mathcal{SV}$.*

Proof. By Propositions 3.8 and 3.14, (1) $\Longrightarrow$ (2) $\Longrightarrow$ (3), for all V. So it remains to prove (3) $\Longrightarrow$ (1). Assume that V is finite-dimensional and in $\mathcal{SV}$. Choose a basis $f_1, \ldots, f_k$ for $L^\infty(V, \mathbb{R})$, and use it to give a smooth linear map $f : V \to \mathbb{R}^k$. Note that $k \leq \dim V$. Since V is in $\mathcal{SV}$, f is injective, and hence surjective. The diffeology on $\mathbb{R}^k$ is the fine diffeology, which is the smallest diffeology making it into a diffeological vector space. The map $f : V \to \mathbb{R}^k$ is a smooth linear bijection, so the diffeology on V must be fine as well (and f must be a diffeomorphism). □

The implication (3) $\Longrightarrow$ (1) also follows from Proposition 3.4, since V in $\mathcal{SV}$ implies that $\dim L(V, \mathbb{R}) \geq \dim L^\infty(V, \mathbb{R}) \geq \dim V = \dim L(V, \mathbb{R})$, and so $L^\infty(V, \mathbb{R}) = L(V, \mathbb{R})$.

Diffeological vector spaces whose finite-dimensional subspaces are fine. Write $\mathcal{FFV}$ for the collection of diffeological vector spaces whose finite-dimensional subspaces are fine. One motivation for studying this collection is the following. In [Christensen and Wu 2016], we defined a diffeology on Hector's tangent spaces [1995] which makes them into diffeological vector spaces. While they are not fine in general, we know of no examples that are not in $\mathcal{FFV}$.

As an example, one can show that $\prod_\omega \mathbb{R}$ is in $\mathcal{FFV}$. This also follows from the next result, which is based on a suggestion of Y. Karshon.

Theorem 3.20. *Every diffeological vector space in $\mathcal{SV}$ is in $\mathcal{FFV}$.*

This result is a special case of Theorem 3.22 below, but we provide a direct proof, since it follows easily from earlier results.

Proof. Let W be a finite-dimensional subspace of V with $V \in \mathcal{SV}$. W is in $\mathcal{SV}$, by Proposition 3.11 (1), and so by Theorem 3.19, W is fine. □

Remark 3.21. (1) Note that it is not in general true that every diffeological vector space in $\mathcal{SV}$ is fine. For example, $\prod_\omega \mathbb{R}$ is in $\mathcal{SV}$ by Remark 3.15 (1), but it is not fine. In fact, [Wu 2015, Example 5.4] showed that there is a countable-dimensional subspace of $\prod_\omega \mathbb{R}$ which is not fine. Incidentally, it follows that $\prod_\omega \mathbb{R}$ is not the colimit in DVect of its finite-dimensional subspaces, since fine diffeological vector spaces are closed under colimits; see [Wu 2015, Property 6 after Definition 5.2].

(2) When $\mathbb{R}$ is equipped with the continuous diffeology (see Example 3.18), it is Hausdorff but is not in $\mathcal{FFV}$. We will see in Proposition 3.23 that the reverse inclusion also fails to hold.

The main result of this section is the following:

Theorem 3.22. *Every diffeological vector space in $\mathcal{SD}$ is in $\mathcal{FFV}$.*

We defer the proof to Section 5.

Furthermore, we have the following result:

Proposition 3.23. *There exists a diffeological vector space which is in $\mathcal{FFV}$ but which is not Hausdorff. In particular, the containment of $\mathcal{SD}$ in $\mathcal{FFV}$ is proper.*

Proof. Let V be the vector space with basis $\mathbb{R}$, and for $r \in \mathbb{R}$ write $[r]$ for the corresponding basis vector of V. Let $f : \mathbb{R} \to \mathbb{R}$ be a bijection such that $f^{-1}(U)$ is dense in $\mathbb{R}$ for every open neighbourhood U of 0 in $\mathbb{R}$. Define $p : \mathbb{R} \to V$ by $p(x) = [f(x)]$ and $\bar{p} : \mathbb{R} \to V$ by $\bar{p}(x) = [x]$. Equip V with the vector space diffeology generated by p and $\bar{p}$. In other words, $q : U \to V$ is a plot if and only if for every $u_0 \in U$ there exist an open neighbourhood U' of u_0 in U and finitely many smooth functions $\alpha_i, \beta_i, \bar{\alpha}_j, \bar{\beta}_j : U' \to \mathbb{R}$ such that for any $u \in U'$,

$$(\dagger) \qquad q(u) = \sum_i \alpha_i(u)\,[f(\beta_i(u))] + \sum_j \bar{\alpha}_j(u)\,[\bar{\beta}_j(u)].$$

(In general, one should include terms with smooth multiples of arbitrary vectors in V, but since $\bar{\beta}_j$ can be constant, this case is included in the above.) First we show that the D-topology on V is not Hausdorff. Suppose that V_0 and V_1 are disjoint D-open subsets of V containing [0] and [1], respectively. Then $U_0 := \bar{p}^{-1}(V_0)$ and $U_1 := \bar{p}^{-1}(V_1)$ are open in $\mathbb{R}$. Since p and $\bar{p}$ are both bijections onto the subset of basis vectors in V, it follows that $p^{-1}(V_0) = f^{-1}(U_0)$ and $p^{-1}(V_1) = f^{-1}(U_1)$. Therefore, $p^{-1}(V_0)$ is dense in $\mathbb{R}$ and so $p^{-1}(V_1)$, which is contained in the complement, must not be open in $\mathbb{R}$, contradicting the assumption that V_1 is D-open. So V is not Hausdorff.

Next we show that V is in $\mathcal{FFV}$. It suffices to show that for any finite subset $A \subseteq \mathbb{R}$, the subspace W spanned by A has the fine diffeology. So let $q : U \to V$ be a plot which lands in W. We must show that for each $a \in A$, the component q_a of q is a smooth function $U \to \mathbb{R}$. This is a local property, so we choose $u_0 \in U$ and express q in the form (†). It suffices to handle each sum in (†) separately, so we begin by assuming that q only has terms involving f. Let $A' = f^{-1}(A)$. By shrinking U' if necessary, we can assume that: (1) for any $b' \in A'$, if $\beta_i(u_0) \neq b'$, then $\beta_i(u) \neq b'$ for all $u \in U'$; and (2) if $\beta_i(u_0) \neq \beta_j(u_0)$, then β_i and β_j have disjoint images. Since f is a bijection, we can rephrase these conditions as: (1') for any $b \in A$, if $f(\beta_i(u_0)) \neq b$, then $f(\beta_i(u)) \neq b$ for all $u \in U'$; and (2') if $f(\beta_i(u_0)) \neq f(\beta_j(u_0))$, then $f \circ \beta_i$ and $f \circ \beta_j$ have disjoint images. Condition (1') implies that for $u \in U'$, $q_a(u)$ is the a-coefficient of

$$\sum_{f(\beta_i(u_0))=a} \alpha_i(u)\,[f(\beta_i(u))].$$

Since $q(u)$ is in W, condition (2') implies that for $r \in \mathbb{R} \setminus A$, we must have

$$(\diamond) \qquad \sum_{f(\beta_i(u_0))=a,\, f(\beta_i(u))=r} \alpha_i(u) = 0.$$

And condition (1') implies that ($\diamond$) also holds for $r \in A \setminus \{a\}$, since the sum is empty in that case. Therefore, $q_a(u)$ can be expressed as

$$\sum_{f(\beta_i(u_0))=a} \alpha_i(u),$$

which is a smooth function of $u \in U'$.

The other sum in (†) is handled in a similar way, replacing f by the identity function throughout.

Finally, Proposition 3.16 implies that V is not in $\mathcal{SD}$, giving the last claim. □

Next we observe that if V is projective (and hence in $\mathcal{SV}$ and $\mathcal{SD}$), it does not follow that all countable-dimensional subspaces of V are fine. We will illustrate this with $V = F(\mathbb{R})$. By [Wu 2015, Corollary 6.4], $F(\mathbb{R})$ is projective.

Proposition 3.24. *Let A be a subset of $\mathbb{R}$, and let V be the subspace of $F(\mathbb{R})$ spanned by A. Then V is fine if and only if A has no accumulation point in $\mathbb{R}$.*

For example, $F(\mathbb{R})$ is not fine. As a more interesting example,

$$A = \{1/n \mid n = 1, 2, \ldots\}$$

spans a countable-dimensional subspace V of $F(\mathbb{R})$ which is not fine. It will follow from Proposition 3.25 that V is not free on any diffeological space.

Proof. ($\Leftarrow$) Let $p : U \to V$ be a plot, where U is open in some $\mathbb{R}^n$. Since V is the span of A, there exist unique functions $h_a : U \to \mathbb{R}$ such that

$$p(x) = \sum_{a \in A} h_a(x)[a].$$

Since A has no accumulation point in $\mathbb{R}$, for each a in A there exists a smooth bump function $\phi_a : \mathbb{R} \to \mathbb{R}$ which takes the value 1 at a and 0 at every other element of A. Associated to ϕ_a is a smooth linear map $\tilde{\phi}_a : F(\mathbb{R}) \to \mathbb{R}$ which sends $[a]$ to 1 and all other basis elements from A to 0. Then $h_a = \tilde{\phi}_a \circ p$, which shows that h_a is smooth.

Next we show that locally p factors through the span of a finite subset of A. Fix $u \in U$. As V is a subspace of $F(\mathbb{R})$, there is an open neighbourhood U' of u in U such that

$$p(x) = \sum_{j=1}^{m} f_j(x)[g_j(x)]$$

for $x \in U'$, where f_j and g_j are smooth functions $U' \to \mathbb{R}$. Shrinking U' if necessary, we can assume that it is contained in a compact subset of U. It follows that the image of each g_j is contained in a compact subset of $\mathbb{R}$ and therefore intersects only finitely many points of A. Since there are only finitely many g_j's, $p|_{U'}$ factors through the span of A' for some finite subset A' of A. That is, $h_a(x) = 0$ for all $x \in U'$ and all $a \in A \setminus A'$.

In summary, identifying the span of A' with $\mathbb{R}^{A'}$, we have factored $p|_{U'}$ as $U' \to \mathbb{R}^{A'} \to V$, where the first map is $x \mapsto (h_a(x))_{a \in A'}$ and the second map sends $f : A' \to \mathbb{R}$ to $\sum_{a \in A'} f(a)[a]$.

($\Rightarrow$) Now we prove that if A has an accumulation point a_0 in $\mathbb{R}$, then V is not fine. Pick a sequence (a_i) in $A \setminus \{a_0\}$ that converges fast to a_0. Choose a smooth function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) \neq 0$ for $1/(2n+1) < x < 1/2n$ for each $n \in \mathbb{Z}^+$, and $f(x) = 0$ for all other x. Choose another smooth function $g : \mathbb{R} \to \mathbb{R}$ such that $g(x) = a_n$ for $1/(2n+1) < x < 1/2n$ for each $n \in \mathbb{Z}^+$, with no constraints on g otherwise. It will necessarily be the case that $g(0) = a_0$, and such a smooth g exists because the sequence was chosen to converge fast. Then the function $p : \mathbb{R} \to V$ defined by $p(x) = f(x)[g(x)]$ is smooth, but there is no open neighbourhood U of 0 so that $p|_U$ factors through a finite-dimensional subspace of V. □

On the other hand, we have:

Proposition 3.25. *Let X be a diffeological space whose underlying set has cardinality less than the cardinality of $\mathbb{R}$. Then the following are equivalent*:

(1) *X is discrete.*

(2) *$F(X)$ is fine.*

(3) *$F(X)$ is projective.*

(4) *$F(X)$ is in $\mathcal{SV}$.*

(5) *$F(X)$ is in $\mathcal{SD}$.*

(6) *$F(X)$ is Hausdorff.*

Proof. That (1) $\Longrightarrow$ (2) is straightforward. The implications (2) $\Longrightarrow$ (3) $\Longrightarrow$ (4) and (5) $\Longrightarrow$ (6) follow from Propositions 3.8, 3.14 and 3.16, while (4) $\Longrightarrow$ (5) is clear. None of these use the assumption on the cardinality of X.

It remains to prove that (6) $\Longrightarrow$ (1). Since the natural injective map $X \to F(X)$ is smooth, it is also continuous when X and $F(X)$ are both equipped with the D-topology. Therefore, X is Hausdorff. We must show that the diffeology on X is discrete. Let $p : U \to X$ be a plot from a connected open subset U of a Euclidean space. We will show that p is constant. If not, then the image of p contains two distinct points $x, x' \in X$ which are connected by a continuous path $q : [0, 1] \to X$. The image of q is compact Hausdorff, and therefore normal. So by Urysohn's lemma, there is a continuous map $l : \mathrm{Im}(q) \to \mathbb{R}$ which separates x and x'. Hence, the image of the composite $l \circ q : [0, 1] \to \mathrm{Im}(q) \to \mathbb{R}$ has cardinality equal to the cardinality of $\mathbb{R}$, which is a contradiction, since $\mathrm{Im}(q) \subseteq X$ has cardinality less than the cardinality of $\mathbb{R}$. □

Part of the above proof is based on the argument in [Hamkins 2015]. Note that we have proved that every Hausdorff diffeological space with cardinality less than the cardinality of $\mathbb{R}$ is discrete. The implication (2) $\Longrightarrow$ (1) is also proved in [Wu 2015, Theorem 5.3], without a constraint on the cardinality of X.

Diffeologies determined by smooth linear functionals.

Definition 3.26. The diffeology on a diffeological vector space V is *determined by its smooth linear functionals* if $p : U \to V$ is a plot if and only if $l \circ p$ is smooth for every $l \in L^\infty(V, \mathbb{R})$. Write $\mathcal{DV}$ for the collection of all such diffeological vector spaces.

Note that any vector space with the indiscrete diffeology is in $\mathcal{DV}$. It follows that being in $\mathcal{DV}$ does not imply any of the other conditions we have studied.

Also note that every diffeological vector space V in $\mathcal{DV}$ is *Frölicher*: $p : U \to V$ is a plot if and only if $f \circ p$ is smooth for every $f \in C^\infty(V, \mathbb{R})$. We do not know if the converse holds.

We will see in Proposition 3.31 that for diffeological vector spaces in $\mathcal{DV}$, the converse of Theorem 3.20 holds. For this, we need the following results.

Lemma 3.27. (1) *If V is in $\mathcal{DV}$ and W is a subspace of V, then W is in $\mathcal{DV}$.*

(2) *Let $\{V_i\}$ be a set of diffeological vector spaces. Then each V_i is in $\mathcal{DV}$ if and only if $\prod V_i$ is in $\mathcal{DV}$.*

Since the category DVect is an additive category, (2) also implies that $\mathcal{DV}$ is closed under taking finite direct sums.

Proof. This is straightforward. □

Proposition 3.28. *Let V be a diffeological vector space. Then V is in $\mathcal{DV}$ if and only if V can be written as a direct sum $V \cong W_0 \oplus W_1$ of diffeological vector spaces, where W_0 is indiscrete and W_1 is in $\mathcal{SV} \cap \mathcal{DV}$.*

Proof. Given V in $\mathcal{DV}$, let W_0 be $\bigcap_{l \in L^\infty(V,\mathbb{R})} \ker(l)$ with the subdiffeology. Since $L^\infty(V, \mathbb{R})$ determines the diffeology on V, W_0 is indiscrete. Let W_1 be the quotient V/W_0, with the quotient diffeology. By Lemma 3.30, we have $V \cong W_0 \oplus W_1$ as diffeological vector spaces. If $v + W_0$ is a nonzero element of W_1, then $v \notin W_0$, so there is a smooth linear functional $l : V \to \mathbb{R}$ such that $l(v) \neq 0$. This l factors through W_1, so it follows that W_1 is in $\mathcal{SV}$. By Lemma 3.27, we know that $W_1 \in \mathcal{DV}$, and hence $W_1 \in \mathcal{SV} \cap \mathcal{DV}$.

The converse follows from Lemma 3.27 and the comment after Definition 3.26. □

Definition 3.29. Following [Wu 2015, Definition 3.15], a diagram

$$0 \longrightarrow W_0 \xrightarrow{i} V \xrightarrow{p} W_1 \longrightarrow 0$$

of diffeological vector spaces is a *short exact sequence of diffeological vector spaces* if it is a short exact sequence of vector spaces, i is an induction, and p is a subduction.

Lemma 3.30. *Let*

$$0 \longrightarrow W_0 \xrightarrow{i} V \xrightarrow{p} W_1 \longrightarrow 0$$

be a short exact sequence of diffeological vector spaces. If W_0 is indiscrete, then the sequence splits smoothly, so that $V \cong W_0 \oplus W_1$ as diffeological vector spaces.

Proof. Let $q : V \to W_0$ be any linear function such that $q \circ i = 1_{W_0}$. Since W_0 is indiscrete, q is smooth. Let $k : V \to V$ be the smooth linear map sending v to $v - i(q(v))$. Then $k \circ i = 0$, so k factors as $j \circ p$, where $j : W_1 \to V$ is smooth and linear. The smooth bijection $V \to W_0 \oplus W_1$ sending v to $(q(v), p(v))$ has a smooth inverse sending (w_0, w_1) to $i(w_0) + j(w_1)$, so the claim follows. □

It follows that many properties of a diffeological vector space are equivalent in this setting:

Proposition 3.31. *Let V be in $\mathcal{DV}$. Then the following are equivalent*:

(1) *V is in $\mathcal{SV}$.*

(2) *V is in $\mathcal{SD}$.*

(3) *V is in $\mathcal{FFV}$.*

(4) *$D(V)$ is Hausdorff.*

(5) *V has no nonzero indiscrete subspace.*

Moreover, being in $\mathcal{DV}$ and satisfying one of these conditions is equivalent to being a subspace of a product of copies of $\mathbb{R}$.

Proof. Without any assumption on V, we have (1) $\Longrightarrow$ (2) $\Longrightarrow$ (3) and (2) $\Longrightarrow$ (4) using Theorem 3.22 and Proposition 3.16. It is easy to see that (3) $\Longrightarrow$ (5) and (4) $\Longrightarrow$ (5). By Proposition 3.28, (5) $\Longrightarrow$ (1) when V is in $\mathcal{DV}$, and so we have shown that the five conditions are equivalent for $V \in \mathcal{DV}$.

For the last claim, a product of copies of $\mathbb{R}$ is in both $\mathcal{SV}$ and $\mathcal{DV}$, and both are closed under taking subspaces. Conversely, if V is in $\mathcal{SV} \cap \mathcal{DV}$, it is easy to check that

$$V \to \prod_{L^\infty(V,\mathbb{R})} \mathbb{R}$$

defined by $v \mapsto (f(v))_{f \in L^\infty(V,\mathbb{R})}$ is a linear induction, and hence V is a subspace of a product of copies of $\mathbb{R}$. □

Remark 3.32. (1) It is not true that every diffeological vector space is in $\mathcal{DV}$. For example, when $\mathbb{R}$ is equipped with the continuous diffeology (see Example 3.18), all smooth linear functionals are zero, but the diffeology is not indiscrete.

(2) Other properties we have studied cannot be added to Proposition 3.31. For example, we saw in Remark 3.15 (1) that $\prod_\omega \mathbb{R}$ is in $\mathcal{SV}$ but is not fine or projective. And it is easy to see that $\prod_\omega \mathbb{R}$ is in $\mathcal{DV}$.

It is not hard to show that every fine diffeological vector space is in $\mathcal{DV}$. As a final example, we will show that not every projective diffeological vector space is in $\mathcal{DV}$, and therefore that none of our other conditions on a diffeological vector space V implies that V is in $\mathcal{DV}$.

We will again use the diffeological vector space $F(\mathbb{R})$, which is projective by [Wu 2015, Corollary 6.4]. We now show that it is not in $\mathcal{DV}$.

Proposition 3.33. *The free diffeological vector space $F(\mathbb{R})$ generated by $\mathbb{R}$ is not in $\mathcal{DV}$.*

Proof. Fix a nonzero smooth function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\operatorname{supp}(\phi) \subset (0, 1)$ and $|\phi(x)| \leq 1$ for all $x \in \mathbb{R}$. For each $n \in \mathbb{Z}^+$, define $\phi_n : \mathbb{R} \to \mathbb{R}$ by

$$\phi_n(x) = \phi\left(\frac{x - \frac{1}{n+1}}{\frac{1}{n} - \frac{1}{n+1}}\right).$$

Finally, define $g : \mathbb{R} \to F(\mathbb{R})$ by

$$g(t) = \begin{cases} 2^{-n}\, \phi_n(t) \sum_{i=1}^{n} \left[\frac{1}{i}\right] & \text{if } \frac{1}{n+1} \leq t < \frac{1}{n}, \text{ for } n > 0, \\ 0 & \text{else.} \end{cases}$$

Then g is not a plot of $F(\mathbb{R})$, since locally around $0 \in \mathbb{R}$, g cannot be written as a finite sum of $f_i(x)[h_i(x)]$, where f_i and h_i are smooth functions with codomain $\mathbb{R}$. But for each $l \in L^\infty(F(\mathbb{R}), \mathbb{R})$,

$$l \circ g(t) = \begin{cases} 2^{-n}\, \phi_n(t) \sum_{i=1}^{n} l\left(\left[\frac{1}{i}\right]\right) & \text{if } \frac{1}{n+1} \leq t < \frac{1}{n}, \\ 0 & \text{else.} \end{cases}$$

This is smooth, since the set $\left\{l\left(\left[\frac{1}{i}\right]\right)\right\}$ is bounded, using the smoothness of l. □

As an easy corollary, we have:

Corollary 3.34. *$F(\mathbb{R})$ is not a subspace of a product of copies of $\mathbb{R}$.*

4. Some applications

Recall that a diagram

$$0 \longrightarrow V_1 \xrightarrow{f} V_2 \xrightarrow{g} V_3 \longrightarrow 0$$

is a short exact sequence of diffeological vector spaces if it is a short exact sequence of vector spaces such that f is an induction and g is a subduction. We say that the sequence *splits smoothly* if there exists a smooth linear map $r : V_2 \to V_1$ such that $r \circ f = 1_{V_1}$, or equivalently, if there exists a smooth linear map $s : V_3 \to V_2$ such that $g \circ s = 1_{V_3}$. In either case, V_2 is smoothly isomorphic to $V_1 \times V_3$; see [Wu 2015, Theorem 3.16].

Not every short exact sequence of diffeological vector spaces splits smoothly. For example, if we write K for the subspace of $C^\infty(\mathbb{R}, \mathbb{R})$ consisting of the smooth functions which are flat at 0, then K is not a smooth direct summand of $C^\infty(\mathbb{R}, \mathbb{R})$ [Wu 2015, Example 4.3].

As a first application of the theory established so far, we can construct additional short exact sequences of diffeological vector spaces which do not split smoothly:

Example 4.1. Let M be a manifold of positive dimension, and let A be a finite subset of M. Write V for the subspace of $F(M)$ spanned by the subset $M \setminus A$ of M. We claim that V is not a smooth direct summand of $F(M)$.

To see this, write W for the quotient diffeological vector space $F(M)/V$. Then, as a vector space, $W = \bigoplus_{a\in A} \mathbb{R}$. So we have a short exact sequence $0 \to V \to F(M) \to W \to 0$ in DVect. Suppose this sequence splits smoothly. By Example 3.7, $F(M)$ is projective, and therefore W is as well. By Proposition 3.14 and Theorem 3.20, W is in $\mathcal{FFV}$. Since W is finite-dimensional, it is fine. But the smooth map $M \to F(M) \to W = \bigoplus_{a\in A} \mathbb{R}$ sends each $a \in A$ to a basis vector and other points in M to 0, so it is not a smooth map in the usual sense. This contradicts the fact that W is fine.

As a second application, we prove:

Theorem 4.2. *Let V be in $\mathcal{SV}$. Then every finite-dimensional subspace of V is a smooth direct summand.*

Proof. Let W be a finite-dimensional subspace of $V \in \mathcal{SV}$. By Theorem 3.20, we know that W has the fine diffeology. Moreover, since V is in $\mathcal{SV}$, there is a smooth linear injective map $V \to \prod_{i\in I} \mathbb{R}$ for some index set I. Since $\prod_{i\in I} \mathbb{R}$ is in $\mathcal{SV}$, again by Theorem 3.20, we know that the composite $W \hookrightarrow V \to \prod_{i\in I} \mathbb{R}$ is an induction, although the second map might not be an induction. So, we are left to prove this statement for the case $V = \prod_{i\in I} \mathbb{R}$.

Write $\dim(W) = m$. By Gaussian elimination, there exist distinct $i_1, \ldots, i_m \in I$ such that the composite $W \hookrightarrow V = \prod_{i\in I} \mathbb{R} \to \mathbb{R}^m$ is an isomorphism of vector spaces, where the second map is the projection onto the $i_1, \ldots, i_m$ coordinates, and hence smooth. Since both W and $\mathbb{R}^m$ have the fine diffeology, this isomorphism is a diffeomorphism, and by composing with its inverse we obtain a smooth linear map $r : V \to W$ such that the composite

$$W \hookrightarrow V \xrightarrow{r} W$$

is 1_W. Therefore, W is a smooth direct summand of V. □

5. Proof of Theorem 3.22

Theorem 3.22. *Every diffeological vector space in $\mathcal{SD}$ is in $\mathcal{FFV}$.*

Proof. If a diffeological vector space is in $\mathcal{SD}$, then so are all of its subspaces. So it suffices to show that every finite-dimensional diffeological vector space in $\mathcal{SD}$ is fine.

Write V for $\mathbb{R}^n$ with the structure of a diffeological vector space which is not fine. We will use the word "smooth" (resp. "continuous") to describe functions $\mathbb{R} \to V$ and $V \to \mathbb{R}$ which are smooth (resp. continuous) with respect to the usual diffeology (resp. topology) on $\mathbb{R}^n$. We use the word "plot" to describe functions $\mathbb{R} \to V$ which are in the diffeology on V, and write $f \in C^\infty(V, \mathbb{R})$ to describe functions which are smooth with respect to this diffeology.

By Proposition 3.3, there is a plot $p : \mathbb{R} \to V$ which is not smooth. Since plots are closed under translation in the domain and codomain, we can assume without loss of generality that $p(0) = 0$ and p is not smooth at $0 \in \mathbb{R}$. We will show that this implies that V is not in $\mathcal{SD}$.

Case 1: Suppose p is continuous at 0. Consider $A := \{\nabla f(x) \mid f \in C^\infty(V, \mathbb{R}), x \in V\}$. Then A is a subset of $\mathbb{R}^n$.

We claim that A is a proper subset of $\mathbb{R}^n$. If A is not proper, then there exist $(f_1, a_1), \ldots, (f_n, a_n) \in C^\infty(V, \mathbb{R}) \times V$ such that $\nabla f_1(a_1), \ldots, \nabla f_n(a_n)$ are linearly independent. Then $g_i : V \to \mathbb{R}$ defined by $g_i(x) = f_i(x + a_i)$ is in $C^\infty(V, \mathbb{R})$, $G := (g_1, \ldots, g_n) : V \to \mathbb{R}^n$ is smooth, and the Jacobian $JG(0)$ is invertible. Therefore, by the inverse function theorem, G is a local diffeomorphism near $0 \in V$. Since $p(0) = 0$, p is continuous at $0 \in \mathbb{R}$, and $G \circ p$ is smooth, it follows that p is smooth at 0, contradicting our assumption on p. So A is a proper subset.

By the same method of translation, one sees that A is a subspace of $\mathbb{R}^n$. Hence, there exists $0 \neq v \in \mathbb{R}^n$ such that $v \perp A$, which implies that $f(x + tv) = f(x)$ for every $f \in C^\infty(V, \mathbb{R})$, $x \in V$ and $t \in \mathbb{R}$, i.e., V is not in $\mathcal{SD}$.

Case 2: Suppose that p is not continuous at 0.

Case 2a: Suppose there exist $k \in \mathbb{N}$ and $\epsilon > 0$ such that $t^k p(t)$ is bounded on $[-\epsilon, \epsilon]$. Let k be the smallest such exponent and write $q(t) := t^k p(t)$, which is also a plot. We claim that q is not smooth at 0. If $k = 0$, then $q = p$, which is assumed to not be smooth at 0. If $k > 0$ and $q'(0)$ exists, then $q(t)/t \to q'(0)$ as $t \to 0$, which implies that $t^{k-1} p(t)$ is also bounded on $[-\epsilon, \epsilon]$, contradicting the minimality of k. So q is not smooth at 0.

If q is continuous at 0, then by Case 1, we are done.

So assume that q is not continuous at 0. Then, since q is bounded on $[-\epsilon, \epsilon]$, there exists a sequence t_i converging to 0 such that $q(t_i)$ converges to a nonzero $v \in V$. If f is in $C^\infty(V, \mathbb{R})$, then $f \circ q$ is smooth, so $f(0) = f(q(0)) = f(q(\lim t_i)) = \lim f(q(t_i)) = f(\lim q(t_i)) = f(v)$. Therefore, the functions in $C^\infty(V, \mathbb{R})$ do not separate points.

Case 2b: Suppose that Case 2a does not apply. Then for each $k \in \mathbb{N}$, $\epsilon > 0$ and $M > 0$, there exists $t \in [-\epsilon, \epsilon]$ such that $\|t^k\, p(t)\| > M$. (Note that $t \neq 0$, since $p(0) = 0$.) Using this for $k = 0$, choose $t_1 \in [-1, 1]$ such that $\|p(t_1)\| > 1$. Then, for each integer $k > 0$, choose t_k with $|t_k| \leq |t_{k-1}|/2$ such that $\|t_k^k\, p(t_k)\| > k$. If $m \leq k$, then t_k also satisfies $\|t_k^m\, p(t_k)\| > k \geq m$, since $|t_k| \leq 1$. Therefore, we can restrict to a subsequence of the t_k all having the same sign. To fix notation, assume that each t_k is positive. Then, for $m \leq k$,

$$\frac{\frac{1}{\|p(t_k)\|}}{t_k^m} < \frac{1}{k}$$

and so, for each m, the left-hand side goes to 0 as $k \to \infty$. By Lemma 5.1, there is a smooth curve $c : \mathbb{R} \to \mathbb{R}$ such that $c(t_k) = 1/\|p(t_k)\|$. It follows that $q(t) := c(t)p(t)$ is a plot such that $q(0) = 0$ and $\|q(t_k)\| = 1$ for each k. Therefore, there is a subsequence converging to a nonzero $v \in V$, and the argument at the end of Case 2a shows that $C^\infty(V, \mathbb{R})$ does not separate points. □

Lemma 5.1 (extended special curve lemma). *Let $\{x_k\}$ and $\{t_k\}$ be sequences in $\mathbb{R}$ such that $0 < t_k < t_{k-1}/2$ for each k and $x_k/t_k^m \to 0$ as $k \to \infty$ for each $m \in \mathbb{Z}^+$. Then there is a smooth function $c : \mathbb{R} \to \mathbb{R}$ such that $c(t_k) = x_k$ for each k and $c(t) = 0$ for $t < 0$.*

The proof closely follows [Kriegl and Michor 1997, page 18], and can easily be generalized further.

Proof. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a smooth function such that $\phi(t) = 0$ for $t \le 0$ and $\phi(t) = 1$ for $t \ge 1$. Define $c : \mathbb{R} \to \mathbb{R}$ by

$$c(t) = \begin{cases} 0 & \text{for } t \le 0, \\ x_{k+1} + \phi\left(\dfrac{t - t_{k+1}}{t_k - t_{k+1}}\right)(x_k - x_{k+1}) & \text{for } t_{k+1} \le t \le t_k, \\ x_1 & \text{for } t_1 \le t. \end{cases}$$

The function c is smooth away from 0 and for $t_{k+1} \le t \le t_k$ we have

$$c^{(r)}(t) = \phi^{(r)}\left(\frac{t - t_{k+1}}{t_k - t_{k+1}}\right)\frac{1}{(t_k - t_{k+1})^r}(x_k - x_{k+1}).$$

Since $t_k - t_{k+1} > t_k/2 > t_{k+1}$, the right-hand side goes to zero as $t \to 0$. Similarly, $c^{(r)}(t)/t \to 0$, which shows that each $c^{(r+1)}(0)$ exists and is 0. So c is smooth. □

We are indebted to Chengjie Yu for the argument used in Case 1 of Theorem 3.22. After we completed Case 2, Yongjie Shi and Chengjie Yu proved this case in more generality in [Shi and Yu 2017].

References

[Christensen and Wu 2016] J. D. Christensen and E. Wu, "Tangent spaces and tangent bundles for diffeological spaces", *Cah. Topol. Géom. Différ. Catég.* **57**:1 (2016), 3–50. MR Zbl

[Christensen et al. 2014] J. D. Christensen, G. Sinnamon, and E. Wu, "The D-topology for diffeological spaces", *Pacific J. Math.* **272**:1 (2014), 87–110. MR Zbl

[Hamkins 2015] J. D. Hamkins, Reply to "Countable path-connected Hausdorff space", MathOverflow, 2015, available at http://mathoverflow.net/q/214537.

[Hector 1995] G. Hector, "Géométrie et topologie des espaces difféologiques", pp. 55–80 in *Analysis and geometry in foliated manifolds* (Santiago de Compostela, Spain, 1994), edited by X. Masa et al., World Sci., River Edge, NJ, 1995. MR Zbl

[Iglesias-Zemmour 2007] P. Iglesias-Zemmour, "Diffeology of the infinite Hopf fibration", pp. 349–393 in *Geometry and topology of manifolds* (Będlewo, Poland, 2005), edited by J. Kubarski et al., Banach Center Publ. **76**, Polish Acad. Sci. Inst. Math., Warsaw, 2007. MR Zbl

[Iglesias-Zemmour 2013] P. Iglesias-Zemmour, *Diffeology*, Math. Surveys and Monographs **185**, Amer. Math. Sci., Providence, RI, 2013. MR Zbl

[Kriegl and Michor 1997] A. Kriegl and P. W. Michor, *The convenient setting of global analysis*, Math. Surveys and Monographs **53**, Amer. Math. Sci., Providence, RI, 1997. MR Zbl

[Shi and Yu 2017] Y. Shi and C. Yu, "Smooth compositions with a nonsmooth inner function", *J. Math. Anal. Appl.* **455**:1 (2017), 52–57. MR Zbl

[Souriau 1984] J.-M. Souriau, "Groupes différentiels de physique mathématique", pp. 73–119 in *South Rhone seminar on geometry, II* (Lyon, 1983), edited by P. Dazord and N. Desolneux-Moulis, Hermann, Paris, 1984. MR Zbl

[Vincent 2008] M. Vincent, *Diffeological differential geometry*, master's thesis, University of Copenhagen, 2008, available at https://tinyurl.com/martinvincent-pdf.

[Wu 2015] E. Wu, "Homological algebra for diffeological vector spaces", *Homology Homotopy Appl.* **17**:1 (2015), 339–376. MR Zbl

Received May 6, 2017. Revised January 29, 2019.

J. DANIEL CHRISTENSEN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WESTERN ONTARIO
LONDON ON
CANADA
jdc@uwo.ca

ENXIN WU
DEPARTMENT OF MATHEMATICS
SHANTOU UNIVERSITY
GUANGDONG
CHINA
exwu@stu.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.93

DEGREE-ONE, MONOTONE SELF-MAPS OF THE PONTRYAGIN SURFACE ARE NEAR-HOMEOMORPHISMS

ROBERT J. DAVERMAN AND THOMAS L. THICKSTUN

We prove that a self-map of the closed Pontryagin surface can be approximated by homeomorphisms if and only if it is monotone and has degree ±1. This adds to a body of theorems, each of which characterizes for some space or class of spaces those self-maps which are approximable by homeomorphisms.

1. Introduction

Given a topological space X, one can ask, "Which surjective self-maps of X are near-homeomorphisms (i.e., approximable by homeomorphisms)?" For X either an n-manifold ($n = 2$ [Daverman 1986, §25], $n = 3$ [Armentrout 1971] (in case X is noncompact, this also depends on the solution to the 3-dimensional Poincaré conjecture), $n = 4$ [Freedman and Quinn 1990], and $n > 4$ [Siebenmann 1972]) or a Hilbert cube manifold [Chapman 1973], the answer is the cell-like self-maps. For X an n-dimensional Menger manifold it is the UV^{n-1}self-maps [Bestvina 1988]. This paper establishes a monotone approximation theorem (Theorem 2.2 here) attesting that, for a (connected) Pontryagin surface P of [Mitchell et al. 1992], the near-homeomorphisms are the self-maps which are monotone and have degree plus or minus one.

Not surprisingly, the proof hinges on a shrinking argument, which appears in Section 10 here. The crucial result toward this end, Corollary 10.5, promises that decompositions induced over finite graphs in the target of the usual type of map are shrinkable. That corollary combines with a homeomorphism extension theorem for maps between Pontryagin disks to complete the proof of the monotone approximation theorem. The section also contains a related theorem for maps between Pontryagin disks that restrict to homeomorphisms between their boundaries.

To set up the shrinking argument a great deal of preliminary effort is required. Most of that effort is directed toward the following intermediate result, called the factor theorem: given a self-map f as in the hypothesis and a locally separating, simple arc A in the Pontryagin surface P, the decomposition space X obtained

MSC2010: primary 57N05, 57P05; secondary 54B15, 54G99, 57N60.

Keywords: Pontryagin surface, Pontryagin disk, monotone map, near-homeomorphism, degree-one map, figure-eight, shrinkability criterion.

from the decomposition of P whose elements are the point-preimages under f of the points in A and singletons in $P - f^{-1}(A)$ is a Pontryagin surface. The factor theorem is stated formally in Section 3; its proof appears at the end of Section 6, based on a related result called the factor reduction theorem. The latter, in turn, is proved in Section 7. Section 8 introduces the notion of a Pontryagin disk, which is a compact subset of a Pontryagin surface that behaves much like a 2-disk in a genuine surface. The main result of the section establishes a controlled equivalence of Pontryagin disks; it has the useful corollary that all homeomorphisms between the boundary curves of Pontryagin disks extend to homeomorphisms between the Pontryagin disks themselves; that result is an essential component of the proof of the monotone approximation theorem. Section 9 introduces the notion of Pontryagin cellularity, a natural analog to the concept of cellularity in 2-manifolds, and a key ingredient in the shrinking arguments.

Pontryagin surfaces and, in particular, Pontryagin disks were introduced by Mitchell, Repovš, and Ščepin [Mitchell et al. 1992], building on a related construction of Pontryagin [1930]. We define Pontryagin surfaces in the next section in a slightly different way than they did, using decompositions into points and figure-eights. Proposition 2.1 attests to the equivalence of this formulation with the original treatment as controlled inverse limits of monotone maps between closed, orientable surfaces. These objects have several interesting features. Connected Pontryagin surfaces are homogeneous [Jakobsche 1991]. A loop L in a locally compact, locally path-connected, locally homologically 1-connected metric space S is null homologous (Borel–Moore homology with $\mathbb{Z}$ coefficients throughout) if and only if it bounds a singular Pontryagin disk in S [Mitchell et al. 1992]. Any map of a Pontryagin disk or Pontryagin surface into a generalized n-manifold, $n > 4$, can be approximated arbitrarily closely by embeddings [Gu 2017].

Any monotone map between closed orientable surfaces must have absolute degree one [Lacher 1977, §7], which might suggest that the degree-one hypothesis in the statement of the monotone approximation theorem is redundant. It is not. However, the construction of the relevant example is quite intricate and the authors will present it in a separate article.

2. Terminology, notation, conventions, and statement of the main result

All maps of spaces will be continuous. A map is *proper* if the preimage of every compact subset of the target is compact. A surjective map is *monotone* if every point preimage is connected. A homotopy f_t of a map $f_0 : X \to Y$ is *supported in* a subset U of Y if, for all $t \in [0, 1]$ and $x \in X - f_0^{-1}(U)$, $f_t(x) = f_0(x)$ and $f_t(f_0^{-1}(U)) \subseteq U$. A map $f : (X, A) \to (Y, B)$ is *split* if $f^{-1}(B) = A$, and a homotopy of such a split map is *admissible* if it is supported in $Y - B$. Given

$f : X \to Y$ and $B \subset Y$, we say f is one-to-one, bijective, onto, etc. over B if $f|f^{-1}(B) : f^{-1}(B) \to B$ is one-to-one, bijective, onto, etc.

A space is *nice* if it is locally compact, locally path-connected, separable, and metrizable (recall that any connected, nice space has an end-point compactification). For a connected, nice space X we denote the one-point compactification, end-point compactification, and space of ends by $\hat{X}$, $\hat{\hat{X}}$, and $e(X)$, respectively (by convention, if X is compact, then $X = \hat{X} = \hat{\hat{X}}$ and $e(X) = \varnothing$). Note that if U is an open, connected subset of a connected, nice space X, then the quotient space $\hat{X}/(\hat{X} - U)$ is homeomorphic to $\hat{U}$. We will often refer to the "quotient-map" $\hat{X} \to \hat{U}$, by which we mean the composition $\hat{X} \to \hat{X}/(\hat{X} - U) \to \hat{U}$ (the maps being those referred to above).

An *exhaustion* of a set X is a sequence $\{X_i\}$ of subsets of X satisfying $X = \bigcup_i X_i$ and, for all i, $X_i \subseteq X_{i+1}$.

We often refer to a collection $\mathcal{E}$ of pairwise-disjoint compacta in a nice space X as a *decomposition* of X. This means the partition of X whose elements are the elements of $\mathcal{E}$ together with all singletons, each of which is contained in no element of $\mathcal{E}$. This partition (or decomposition) is upper semicontinuous (and hence, the associated decomposition space is metrizable) whenever the elements of $\mathcal{E}$ form a null sequence with respect to some metric on X. Such a decomposition space will be denoted as $X/\mathcal{E}$.

Any space homeomorphic to the wedge of two circles is a *figure-eight*.

Definition. A connected, nice space P is a *Pontryagin surface* if there exists a countable family $\mathcal{E}$ of pairwise-disjoint figure-eights in P such that $\mathcal{E}$ is null in $\hat{P}$ and, for any cofinite subfamily $\mathcal{D}$ of $\mathcal{E}$, the image of P under the decomposition map $\hat{P} \to \hat{P}/\mathcal{D}$ is an orientable surface without boundary. Such a family $\mathcal{E}$ is a *sufficient family* for P. (Observe that any closed orientable surface Q is a Pontryagin surface and that any finite family of pairwise-disjoint figure-eights in Q is a sufficient family if and only if it satisfies the following condition: for any element e of the family, the quotient space Q/e is a surface). If, in addition, the image of P in $\hat{P}/\mathcal{E}$ is either planar or a 2-sphere, then $\mathcal{E}$ is a *full* family for P. A nice space is a Pontryagin surface if each of its components is a Pontryagin surface, and a family $\mathcal{E}$ of figure-eights in a Pontryagin surface is a *sufficient family* if the elements of $\mathcal{E}$ in each component Y constitute a sufficient family for Y. A Pontryagin surface is *closed* if it is compact; otherwise, it is *open*.

A subspace C of a Pontryagin surface X is $\mathbb{P}$-*negligible* if X has a sufficient family no element of which meets C. The *manifold set of* X, denoted $M(X)$, is $\{p \in X \mid p \text{ has a neighborhood homeomorphic to } \mathbb{R}^2\}$. X is *rich* if $M(X) = \varnothing$. It should be noted that, unlike [Mitchell et al. 1992], in order to promote greater generality and to accommodate some of our constructions, we do not assume all Pontryagin surfaces have empty manifold set.

A compact space $\mathbb{D}$ is a *Pontryagin disk* if it is homeomorphic to the closure of some complementary component of a separating simple closed curve in a rich, connected Pontryagin surface. (It is important to keep in mind that Pontryagin disks, unlike Pontryagin surfaces, never contain open 2-disks.) Note that, by Corollary 3.2, the frontier of a Pontryagin disk in a Pontryagin surface is $\mathbb{P}$-negligible.

A compact 1-manifold A in a space X is *locally separating* if, given $p \in A - \partial A$ and any neighborhood U of p, there exists a connected neighborhood V of p such that $V \subset U$, $V \cap A$ is connected, and $V - A$ is not connected.

The Čech n-homology with G coefficients of a compact, metrizable space X will be denoted $\check{H}_n(X; G)$ (however, if $G = \mathbb{Z}$, the coefficient group will not be indicated).

A map $f : X \to Y$ of compact, metrizable spaces is an *$\check{H}_2$-isomorphism, monomorphism, etc.* if it induces an isomorphism, monomorphism, etc. on Čech 2-homology.

A surjective map of closed, orientable surfaces is *standard* if it is bijective over the complement of a finite subset F of the target and the preimage of each point in F is a figure-eight.

A map $f : X \to Y$ of one compact, metrizable space onto another is a *near-homeomorphism* if, given a metric ρ on Y and $\varepsilon > 0$, there exists a homeomorphism $h : X \to Y$ such that, for all $x \in X$, $\rho(f(x), h(x)) < \varepsilon$.

The following proposition provides, in effect, an alternate definition of "closed, connected Pontryagin surface".

Proposition 2.1. *A space X is a closed, connected Pontryagin surface if and only if it is the inverse limit of a sequence $\{p_n : R_{n+1} \to R_n\}_{n=1}^{\infty}$ of standard maps between closed, connected, orientable surfaces such that if, for each $n \in \mathbb{N}$, F_n denotes the finite subset of R_n referred to in the definition of standard map, then, for all $n \neq m$, $p_{n,1}(F_n) \cap F_1 = p_{m,1}(F_m) \cap p_{n,1}(F_n) = \varnothing$ (where $p_{n,1} = p_1 \circ p_2 \circ \cdots \circ p_{n-1}$).*

Proof. (Only if) Suppose $\mathcal{E} = \{e_1, e_2, \ldots\}$ is a sufficient family for X. Set $\mathcal{E}_n = \{e_n, e_{n+1}, \ldots\}$ and form the decomposition space $R_n = X/\mathcal{E}_n$. Note that $R_n = R_{n+1}/e_n$. Let $p_n : R_{n+1} \to R_n$ be the obvious map.

(If) Clearly

$$\mathcal{E} = \{e \subset X \mid \text{there exist } n \in \mathbb{N} \text{ and } x \in F_n \text{ such that } p_{\infty,n}^{-1}(x) = e\}$$

(where $p_{\infty,n}$ denotes the projection of X to R_n) is a countable, pairwise-disjoint, null family of figure-eights in X. To verify that it is sufficient, let $\mathcal{E}'$ be a cofinite subfamily of $\mathcal{E}$ and denote $\mathcal{F} = \mathcal{E} - \mathcal{E}'$ and $\mathcal{F}_n = \{e \in \mathcal{E} \mid e \text{ is a component of } p_{\infty,n}^{-1}(F_n)\}$. Choose $N \in \mathbb{N}$ such that $\mathcal{F} \subset \bigcup_{n=1}^{N} \mathcal{F}_n$. The decomposition space $X/\bigcup_{n=N+1}^{\infty} \mathcal{F}_n$ is R_N whose decomposition space obtained from the decomposition $\{p_{\infty,N}(e) \mid e \in \bigcup_{n=1}^{N} \mathcal{F}_n\}$ is R_1. Apply the parenthetical observation in the above definition of Pontryagin surface to complete the proof. □

Definition. We have from Proposition 2.1 that if P is a closed, connected Pontryagin surface, then $\check{H}_2(P) = \mathbb{Z}$ and, for $m \in \mathbb{N}$, $\check{H}_2(P; \mathbb{Z}_m) = \mathbb{Z}_m$. Given a map $f : P \to Q$ of closed, connected Pontryagin surfaces and choices $\mathcal{O}_P$ and $\mathcal{O}_Q$ of generators of $\check{H}_2(P)$ and $\check{H}_2(Q)$ (but $\mathcal{O}_P = \mathcal{O}_Q$ if $P = Q$), the *degree of* f is the integer n such that the induced homomorphism on Čech 2-homology sends $\mathcal{O}_P$ to $n\mathcal{O}_Q$. Note that the absolute value of the degree is independent of the choice of generators. Our interest focuses on maps of degree one, by which we really mean maps of absolute degree one.

Theorem 2.2 (monotone approximation theorem). *A map $f : P \to Q$ of closed, connected, rich Pontryagin surfaces is a near-homeomorphism if and only if it is monotone and has (absolute) degree one.*

Proof of "only if". That a near homeomorphism must be monotone follows from the well known result [Kuratowski and Lacher 1969] that any uniformly convergent sequence of monotone maps between compact, locally connected metric spaces has a monotone limit. To show that f must have degree one, let $p : Q \to S^2$ be a map arising as the inverse limit of standard maps between closed orientable surfaces and let $f = \lim_{i\to\infty} h_i$ where the $\{h_i : P \to Q\}$ are homeomorphisms. Then $p \circ f = \lim_{i\to\infty}(p \circ h_i)$.

Since S^2 is an ANR there exists an integer $k > 0$ such that $p \circ h_k$ is homotopic to $p \circ f$. Hence, $\deg(p \circ f) = \deg(p) \cdot \deg(f) = \deg(p) \cdot \deg(h_k) = 1$. So $\deg(f) = 1$. □

Applying the Vietoris–Begle mapping theorem, we obtain:

Corollary 2.3. *All cell-like maps between closed, connected, rich Pontryagin surfaces are near-homeomorphisms.*

We adopt the following notational conventions. If A is a subset of a topological space X, then $\operatorname{Fr} A$ and $\operatorname{Int} A$ will denote the frontier and interior of A in X. If A is a manifold, then $\mathring{A}$ denotes $A - \partial A$. $I = [-1, +1] \subset \mathbb{R}$ (but we will also consider I to be the set $[-1, +1] \times \{0\}$ in $\mathbb{R}^2$). $S^1 =$ the unit circle in $\mathbb{R}^2$. $S^2 =$ the one-point compactification of $\mathbb{R}^2$ (so we can regard $\mathbb{R}^2$ as a subset of S^2). $H = \{(x, y) \in \mathbb{R}^2 \mid y \geq 0\}$.

Remarks. Existence, uniqueness up to homeomorphism, and homogeneity of the connected, closed, rich Pontryagin surface (denoted by P in these remarks) are well known. However, it is worth noting that existence follows easily from Proposition 2.1 while uniqueness and homogeneity follow from Corollary 8.2 of this paper (we leave this as an exercise). Note also that any self-map of P constructed as follows is monotone and degree-one but not cell-like. Let $\mathcal{E}$ be a sufficient family for P and $\mathcal{F} \subset \mathcal{E}$ such that the image of $\mathcal{E} - \mathcal{F}$ is dense in $P/\mathcal{E}$ (e.g., $\mathcal{F}$ is finite). It follows that $P/\mathcal{F}$ is a rich Pontryagin surface, so the composition $P \xrightarrow{d} P/\mathcal{F} \xrightarrow{h} P$ where d is the decomposition map and h a homeomorphism is the desired map.

Let S be a compact metric space, G an upper semicontinuous decomposition of S, and $\pi : S \to S/G$ the decomposition map. Then G is *shrinkable* if the following condition, called the Bing shrinkability criterion, is satisfied: for each $\varepsilon > 0$ there exists a homeomorphism $h : S \to S$ such that each $h(g)$ $(g \in G)$ has diameter less than ε, and π and πh are ε-close.

The notion of shrinkability was introduced by R. H. Bing [1952]. He exploited it to provide an effective general method for determining the topological type of certain decomposition spaces. R. D. Edwards [1980] gave an elegant proof for the crucial compact case mentioned below; his proof also can be found in [Daverman 1986, Lemma 6.1].

Theorem 2.4. *An upper semicontinuous decomposition G of a compact metric space S is shrinkable if and only if the decomposition map $\pi : S \to S/G$ is a near-homeomorphism.*

Another setting in which upper semicontinuous decompositions arise involves a proper map $f : X \to Y$ defined on a nice space X and a subset C of Y. The decomposition $G(C)$ of X *induced over* C is the partition consisting of the sets $\{f^{-1}(c) \mid c \in C\}$ and the singletons from $X - f^{-1}(C)$. Here $G(C)$ is upper semicontinuous (and $X/G(C)$ is metrizable) whenever C is closed in Y.

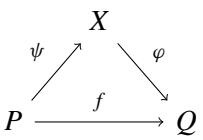
Corollary 2.5. *Let $f : X \to Y$ be a surjective mapping between compact metric spaces and C a closed subset of Y. If the decomposition $G(C)$ induced over C is shrinkable, then f can be approximated, arbitrarily closely, by a surjective map that is injective over C.*

Proof. If $\theta : X \to X/G(C)$ is a homeomorphism very close to the decomposition map $\pi : X \to X/G(C)$, then $F = f\pi^{-1}\theta$ is a map close to f which is 1-1 over C. □

3. The factor theorem

The theorem stated below is a key technical ingredient in the proof of the approximation theorem. Its proof occupies the following four sections.

Theorem 3.1 (the factor theorem). *Suppose the commutative diagram*



of maps and spaces satisfies the following conditions:

(1) *P and Q are closed, connected Pontryagin surfaces.*

(2) *All maps are surjective and f is both monotone and degree one.*

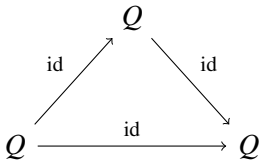
(3) *There is a subspace A of Q which is either a locally separating simple arc or a separating simple closed curve such that φ is injective over A and ψ is injective over $X - \varphi^{-1}(A)$.*

Then X is a Pontryagin surface and $\varphi^{-1}(A)$ is $\mathbb{P}$-negligible in X.

Note. We will prove the factor theorem in detail only for the case in which A is an arc. The proof for A a simple closed curve is essentially the same except for some minor details which we leave to the reader.

Corollary 3.2. *Any locally separating arc or separating simple closed curve in a closed Pontryagin surface is $\mathbb{P}$-negligible.*

Proof. Let Q be the closed Pontryagin surface and apply the factor theorem to the diagram



□

4. Sufficient families

In this brief section we state and prove some results and their consequences concerning sufficient families. First we present some definitions and notation.

A family $\mathcal{D}$ of compacta in a locally compact space X is *locally finite* if, for any compact subset C of X, the set $\{e \in \mathcal{D} \mid e \cap C \neq \varnothing\}$ is finite. If $\mathcal{E}$ is a family of compacta in X and $\mathcal{D} \subset \mathcal{E}$, we say $\mathcal{D}$ is a *locally cofinite* subfamily of $\mathcal{E}$ if $\mathcal{E} - \mathcal{D}$ is locally finite. We denote, for a subset U of X, $\mathcal{E}(U) = \{e \in \mathcal{E} \mid e \subset U\}$.

Observation. *Any locally cofinite subfamily of a sufficient family is sufficient.*

Proposition 4.1. *Any open subset U of a Pontryagin surface P is a Pontryagin surface. Furthermore, if $\mathcal{E}$ is a sufficient family for P, then $\mathcal{E}(U)$ is a sufficient family for U.*

Proof. We can assume that U is connected. Let $\mathcal{D}$ be a locally cofinite subfamily of $\mathcal{E}(U)$. It will suffice to show that if V is an open, connected subset of U such that $\overline{V} \subset U$ and $\overline{V}$ is compact, then the image of V under the decomposition map $\hat{P} \to \hat{P}/\mathcal{D}$ is a surface. Denote

$$\mathcal{F} = \{e \in \mathcal{E} \mid e \text{ meets both } P - U \text{ and } \overline{V} \text{ or } e \subset V \text{ and } e \notin \mathcal{D}\}.$$

Note that $\mathcal{F}$ is finite and $\mathcal{D} \subset \mathcal{E} - \mathcal{F}$. The image of V in $\hat{P}/\mathcal{D}$ is sent homeomorphically by the obvious decomposition map onto its image in $\hat{P}/(\mathcal{E} - \mathcal{F})$, which must be a surface since $\mathcal{E} - \mathcal{F}$ is sufficient for P. □

Lemma 4.2. *Suppose U is a Pontryagin surface with sufficient family $\mathcal{D}$. If A is a closed subset of U with $A \subset M(U)$, then the family $\{e \in \mathcal{D} \mid e \cap A = \varnothing\}$ is sufficient for U.*

Proof. Verify that $\{e \in \mathcal{D} \mid e \cap A \neq \varnothing\}$ is locally finite and apply the above observation. □

Proposition 4.3. *A connected, nice space U is a Pontryagin surface if and only if $\hat{\hat{U}}$ is a Pontryagin surface.*

Proof. (If) This follows from the previous proposition.

(Only if) Let $\mathcal{E}$ be a sufficient family for U and let R denote the image of U under the decomposition map $d : \hat{\hat{U}} \to \hat{\hat{U}}/\mathcal{E}$. So R is an open, connected, orientable surface whose end-point compactification is $\hat{\hat{U}}/\mathcal{E}$. From the classification theorem for open surfaces [Richards 1963] we obtain a locally finite (in R) family $\mathcal{D}$ of figure-eights in R such that $R/\mathcal{D}$ is planar and R is null in $\hat{\hat{U}}/\mathcal{E}$. By a standard general position argument we can choose these figure-eights to avoid $\bigcup_{e\in\mathcal{E}} d(e)$. Then $\mathcal{E} \cup \{d^{-1}(e) \mid e \in \mathcal{D}\}$ is a sufficient family for $\hat{\hat{U}}$. □

5. 2-coherence

This section includes a series of lemmas, propositions, and theorems to be used in the proof of the factor theorem, most of which take as their hypotheses only certain Čech-homological properties of Pontryagin surfaces. Any nice space having these properties will be termed *2-coherent*.

Definition. Suppose X is a connected, nice space and $\check{H}_2(\hat{X}) = \mathbb{Z}$. A family $\mathcal{U}$ of open, connected, nonempty subsets of X is a *coherence family* if, for any $U \in \mathcal{U}$, the following conditions are satisfied:

(1) The quotient map $\hat{X} \to \hat{U}$ is an $\check{H}_2$-isomorphism.

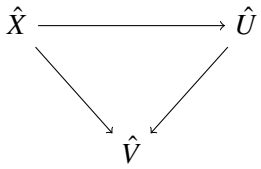
(2) For $n \in \mathbb{N} - \{1\}$, $\check{H}_2(\hat{U}; \mathbb{Z}_n) = \mathbb{Z}_n$ and the homomorphism $\check{H}_2(\hat{U}) \to \check{H}_2(\hat{U}; \mathbb{Z}_n)$ (induced by the coefficient group epimorphism $\mathbb{Z} \to \mathbb{Z}_n$) is an epimorphism.

(3) Any open, connected, nonempty subset of X is exhausted by elements of $\mathcal{U}$.

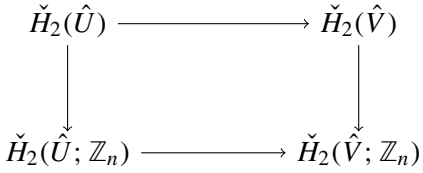
Definition. A connected, nice space X with $\check{H}_2(\hat{X}) = \mathbb{Z}$ is *2-coherent* if the class of all open, connected, nonempty subsets of X is a coherence family. Among its other benefits, 2-coherence characterizes the 2-manifolds within the class of 2-complexes.

A proper map $f : X \to Y$ of 2-coherent spaces has (*absolute*) *degree one* if $\hat{f} : \hat{X} \to \hat{Y}$ induces an isomorphism on Čech 2-homology with $\mathbb{Z}$ coefficients.

Observation. *If $\mathcal{U}$ is a coherence family for X and $V, U \in \mathcal{U}$ with $V \subset U$, then the quotient map $\hat{U} \to \hat{V}$ is an $\check{H}_2$-isomorphism. (To see this, apply $\check{H}_2$ to the diagram*

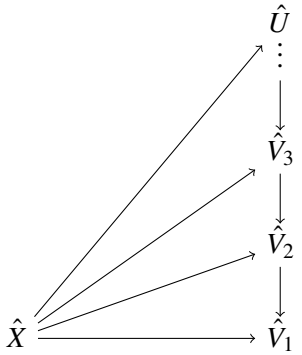


where the maps are quotient maps.) Moreover, for $n \in \mathbb{N} - \{1\}$, $\check{H}_2(\hat{U}; \mathbb{Z}_n) \to \check{H}_2(\hat{V}; \mathbb{Z}_n)$ is an isomorphism. To see this consider the commutative diagram



Lemma 5.1. *If a connected nice space X with $\check{H}_2(\hat{X}) = \mathbb{Z}$ has a coherence family, then it is 2-coherent.*

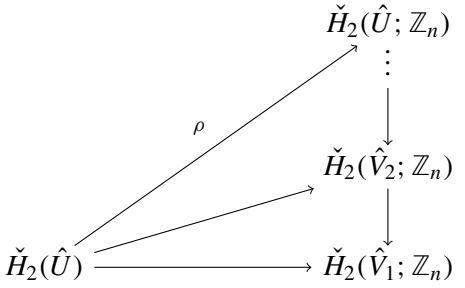
Proof. Let $\mathcal{V}$ denote the coherence family and let U be an open, connected subset of X. Let $\{V_i\}_{i=1}^{\infty}$ be an exhaustion of U with $V_i \in \mathcal{V}$ for all i. To verify that (1) (in the definition of coherence family) holds for U, apply the continuity axiom for Čech homology to the diagram obtained by applying $\check{H}_2$ to the following commutative diagram of spaces and maps:



All maps are quotient maps. Note that $\hat{U} = \varprojlim(\hat{V}_1 \leftarrow \hat{V}_2 \leftarrow \cdots)$.

One obtains from the same diagram that $\check{H}_2(\hat{U}; \mathbb{Z}_n) = \mathbb{Z}_n$. To verify that $\check{H}_2(\hat{U}) \to \check{H}_2(\hat{U}; \mathbb{Z}_n)$ is onto first note that, for all i, the composition $\check{H}_2(\hat{U}) \to \check{H}_2(\hat{V}_i) \to \check{H}_2(\hat{V}_i; \mathbb{Z}_n)$ is onto (the first homomorphism is an isomorphism and the

second is onto by hypothesis). Now consider the commutative diagram



The "vertical" maps are isomorphisms and the others (with the possible exception of ρ) are onto. Hence, ρ is onto. □

The proof of the following lemma is left to the reader.

Lemma 5.2. *If C is a closed 0-dimensional subset of a compact, metrizable space X, then the quotient map $X \to X/C$ is an $\check{H}_2$-isomorphism.*

Proposition 5.3. *A connected nice space U is 2-coherent if and only if $\hat{\hat{U}}$ is 2-coherent.*

Proof. (If) This part is left to the reader.

(Only if) We claim that

$$\mathcal{V} = \{V \subset \hat{\hat{U}} \mid V \text{ is connected and open, and } V \cap e(U) \text{ is compact}\}$$

is a coherence family. We verify only condition (1) in the definition of coherence family and leave the rest to the reader. Consider the commutative diagram

$$\begin{array}{ccc} \hat{\hat{U}} & \longrightarrow & \hat{\hat{U}}/(\hat{\hat{U}} - V) = \hat{V} \\ \downarrow & & \downarrow \\ \hat{U} & \longrightarrow & \hat{V} \end{array}$$

where $V \in \mathcal{V}$ and the maps are the obvious quotient maps (the "vertical" map on the right sends $e(U) \cap V$ to $\hat{V} - V$). Applying $\check{H}_2$ to the diagram we have, by hypothesis, that the bottom horizontal homomorphism is an isomorphism and the vertical homomorphisms are isomorphisms by Lemma 5.2, so the top horizontal homomorphism is an isomorphism. □

Proposition 5.4. *Every connected Pontryagin surface is 2-coherent.*

Proof. Observe first that since the end-point compactification of a Pontryagin surface is a Pontryagin surface (Proposition 4.3) and any open connected subset of a 2-coherent space is 2-coherent, we can assume without loss of generality that the Pontryagin surface P of the hypothesis is compact. Use Proposition 2.1 to express P as the inverse limit of standard maps $\{p_n : R_{n+1} \to R_n\}_{n=1}^{\infty}$. We leave it to the reader to verify that the following class of open sets is a coherence family for P:

$\big\{V \mid$ there exists a connected compact subsurface M_n of R_n such that $p_{\infty,n}$ is one-to-one over ∂M_n and V is the interior of $p_{\infty,n}^{-1}(M_n)\big\}$. □

Lemma 5.5. *The proper cell-like image of a* 2*-coherent space is* 2*-coherent.*

Proof. Apply the Vietoris–Begle theorem. □

Lemma 5.6. *Suppose X is a connected* 2*-coherent space. Then*:

(1) *X contains no locally separating point.*

(2) *X contains no separating, closed* 0*-dimensional subset.*

(3) *X contains no separating set which is the union of a simple arc and a closed* 0*-dimensional set.*

(4) *If X is separated by a set which is the union of a simple closed curve α and a closed* 0*-dimensional set, then α separates X.*

Proof. (1) Suppose U is a connected open set in X and $p \in U$ such that $U - \{p\}$ is not connected. Since U is 2-coherent we have that $\hat{\hat{U}}$ is compact, 2-coherent, and separated by p. Denote by C the closure of a component of $\hat{\hat{U}} - \{p\}$ and let D be the closure of the union of all other components of $\hat{\hat{U}} - \{p\}$. Then $\check{H}_2(\hat{\hat{U}}) = \check{H}_2(C) \oplus \check{H}_2(D)$ and so one of the two summands must be trivial, which is impossible by the 2-coherence of $\hat{\hat{U}}$.

(2) The proof is similar to that of (1).

(3) Suppose otherwise and let A denote the arc. By Lemma 5.5, X/A is 2-coherent and is separated by a closed 0-dimensional set, which contradicts (2).

(4) Since $\hat{\hat{X}}$ is 2-coherent by Proposition 5.3 we can assume without loss that X is compact. Denote the 0-dimensional set by C and suppose α does not separate X. Denote $U = X - \alpha$ and note that, by connectivity of U, $\hat{U}$ is 2-coherent. Since X/α is the one-point compactification of $X - \alpha$ we have the natural map $\varphi : \hat{\hat{U}} \to X/\alpha$ (from the end-point compactification of any nice space to the one-point compactification of that space). Let x denote the image of α under the quotient map $X \to X/\alpha$ and note that the map $\varphi : (\hat{\hat{U}}, e(U) \cup C) \to (X/\alpha, \{x\} \cup C)$ is split (by abuse of notation C is considered to be a subset of both $\hat{\hat{U}}$ and X/α). However, $e(U) \cup C$ cannot separate $\hat{\hat{U}}$ by (2) and hence $\{x\} \cup C$ cannot separate X/α. But then $\alpha \cup C$ cannot separate X. □

Corollary 5.7. *If X is compact and 2-coherent and A is a cell-like subset of X, then $X - A$ has one end.*

Proof. Otherwise, A would be a locally separating point in X/A (which is 2-coherent by Lemma 5.5). □

Corollary 5.8. *Suppose A is a separating simple closed curve in a compact 2-coherent space X.*

(1) *If U is any component of $X - A$, then $\overline{U} = U \cup A$.*

(2) *A is locally separating.*

Proof. (1) Suppose $A \not\subset \overline{U}$. Then $\overline{U} \cap A$ is contained in some arc α in A which would make α a separating point in the quotient space X/α (which would be 2-coherent).

(2) Use (1). □

Lemma 5.9. *Suppose X is a compact metrizable space, S is a simple closed curve in X, and A and B are the closures in X of two distinct components of $X - S$. Then the inclusion-induced homomorphism $\check{H}_2(A) \oplus \check{H}_2(B) \to \check{H}_2(X)$ is injective.*

Proof. There exists a sequence of nerves $\{p_{n+1,n} : (X_{n+1}, A_{n+1}, B_{n+1}, S_{n+1}) \to (X_n, A_n, B_n, S_n)\}_{n=1}^{\infty}$ such that, for each n, S_n is a simple closed curve, A_n and B_n are closed components of $X_n - S_n$, and for $Z = X$, A, B, or S, $\varprojlim\{p_{n+1,n} : Z_{n+1} \to Z_n\}$ is Z. Now conclude from a Mayer–Vietoris sequence that, for each n, $H_2(A_n) \oplus H_2(B_n) \to H_2(X_n)$ is injective. Since the inverse limit of monomorphisms is a monomorphism, the conclusion follows. □

Observation. *If E is a compact subspace of a 2-coherent space X with $E \neq X$, then the inclusion-induced homomorphism $\check{H}_2(E; G) \to \check{H}_2(\hat{X}; G)$ (where $G = \mathbb{Z}$ or $\mathbb{Z}_n$ for some $n \in \mathbb{N} - \{1\}$) is trivial.*

Proof. Let U be a component of $X - E$ and note that the composition $\check{H}_2(E; G) \to \check{H}_2(\hat{X}; G) \to \check{H}_2(\hat{U}; G)$ (induced by the obvious maps $E \to \hat{X} \to \hat{U}$) is trivial and the second of the two homomorphisms is an isomorphism. □

Lemma 5.10. *Suppose S is a separating simple closed curve in a compact 2-coherent space X and U is a component of $X - S$. Then*:

(1) *$\check{H}_2(\overline{U}) = 0$.*

(2) *$\partial_* : \check{H}_2(\overline{U}, S) \to \check{H}_1(S)$ is an isomorphism.*

(3) *$X - S$ has two components.*

Proof. (1) By Lemma 5.9 (where $A = \overline{U}$) we have $\check{H}_2(\overline{U}) \to \check{H}_2(X)$ is injective. But, by the observation, it is also trivial.

(2) We have homomorphisms $\check{H}_2(\hat{U}) \xrightarrow{\varphi} \check{H}_2(\overline{U}, S) \xrightarrow{\partial_*} \check{H}_1(S)$ where φ is the inverse of the isomorphism induced by the quotient map $\overline{U} \to \overline{U}/S = \hat{U}$. Let $\alpha \in \check{H}_2(\hat{U})$

be a generator and denote $\beta = (\partial_* \circ \varphi)(\alpha)$. We will show that β is a generator of $H_1(S)$ (which we identify with $\mathbb{Z}$). We can assume without loss of generality that $\beta \geq 0$. If $\beta = 0$, then $\varphi(\alpha)$ is in the image of $\check{H}_2(\overline{U}) \to \check{H}_2(\overline{U}, S)$ and hence $\check{H}_2(\overline{U})$ is nontrivial (impossible by (1)). If $\beta > 1$, we have a nontrivial element of $\check{H}_2(\overline{U}; \mathbb{Z}_\beta)$ again violating (1) (note that nontriviality of the element follows from the surjectivity of $\check{H}_2(\hat{U}) \to \check{H}_2(\hat{U}; \mathbb{Z}_\beta)$).

(3) Assume $X - S$ has at least two components U_1 and U_2. If $(X - S) \neq U_1 \cup U_2$, use (2) to argue that $\check{H}_2(\overline{U}_1 \cup \overline{U}_2)$ must be nontrivial, contradicting (1). □

Corollary 5.11. *If X is a noncompact 2-coherent space and R is a separating closed subset of X homeomorphic to $\mathbb{R}$, then $X - R$ has two components.*

Proof. Let α denote the closure in $\hat{\hat{X}}$ of R. By Lemma 5.10 it will suffice to show that α is a simple closed curve, but if α were an arc, then the 2-coherent space $\hat{\hat{X}}/\alpha$ would contain a separating point. □

Definition. A simple arc A in a 2-coherent space X is *2-sided* if, given any subarc α of A, there exists a neighborhood V such that $V \cap A = \mathring{\alpha}$, $V - A$ has two components, and denoting the two components by V_1 and V_2, $\overline{V}_1 \cap \overline{V}_2 = \alpha$ and $\mathrm{Fr}(\overline{V}_i) = \overline{V}_i - V_i$ $(i = 1, 2)$. (Such a neighborhood V of $\mathring{\alpha}$ will be called *dichotomous*.)

Proposition 5.12. *Any locally separating arc A in a 2-coherent space X is 2-sided.*

Proof. By Corollary 5.11 it will suffice to show that $\mathring{A}$ separates some open connected neighborhood of itself. We briefly outline the proof. Construct a family $\mathcal{U} = \{U_n\}_{n \in \mathbb{Z}}$ of open connected sets covering $\mathring{A}$ and satisfying the following properties:

(1) For each n, $U_n - A$ is disconnected.

(2) For each n, $U_n \cap A$ is an open subarc of A whose closure in A is disjoint from ∂A.

(3) $U_i \cap U_j \neq \varnothing$ if and only if $|i - j| \leq 1$.

Now, by a Lebesgue number argument applied infinitely many times, we can choose a second covering $\{V_m\}_{m \in \mathbb{Z}}$ satisfying the same three properties and, in addition, for $|i - j| = 1$, $V_i \cup V_j$ is contained in some element of $\mathcal{U}$. Prove by induction on $N \in \mathbb{N}$ that $\bigcup_{m=-N}^{N} V_m$ is separated by A. Then $\bigcup_{m \in \mathbb{Z}} V_m$ is the desired neighborhood. □

Observation. *If A is a 2-sided simple arc in a 2-coherent space X, then any subarc of A is also 2-sided. Also note that if U is a dichotomous neighborhood of $\mathring{A}$ and V is a connected, open set with $V \subset U$ such that $V \cap A$ is connected, then V is a dichotomous neighborhood of $V \cap A$.*

The following observation is used in the proof of Proposition 5.13. Its proof is left to the reader.

Observation. *The absolute degree of a map of compact* 2*-coherent spaces is "determined locally"*; *i.e.*, *if* $f : X \to Y$ *is such a map and* V *is an open, nonnull, connected subset of* Y *such that* $f^{-1}(V)$ *is connected, then the absolute degree of* f *is the same as the absolute degree of the one-point compactification of the map* $f| : f^{-1}(V) \to V$.

Proposition 5.13 (degree-one proposition). *If* X *is a compact* 2*-coherent space, then a map* $f : X \to S^2$ *has degree one if the following conditions are satisfied*:

(1) $f^{-1}(S^1)$ *is the union of a simple closed curve* A *and a closed* 0*-dimensional set and* $f|A : A \to S^1$ *is bijective.*

(2) *For* C *either component of* $S^2 - S^1$, $f^{-1}(C) \neq \varnothing$.

Proof. First note that since $f^{-1}(S^1)$ separates X we have, by Lemma 5.6, that A separates X and hence, by Lemma 5.10, that $X - A$ has two components. Let U be one of them and let D be that component of $S^2 - S^1$ which contains $f(U)$ (and hence by condition (2) we have $f^{-1}(D) = U$). Consider the commutative diagram

$$\begin{array}{ccc} \check{H}_2(\overline{U}, A) & \xrightarrow{\partial_*} & \check{H}_1(A) \\ \downarrow{\scriptstyle (f|\overline{U})_*} & & \downarrow{\scriptstyle (f|A)_*} \\ \check{H}_2(\overline{D}, S^1) & \xrightarrow{\partial_*} & \check{H}_1(S^1) \end{array}$$

By Lemma 5.10, $(f|A)_* \circ \partial_*$ is an isomorphism and ∂_* at the bottom of the diagram is obviously an isomorphism. Hence, $(f|\overline{U})_*$ is an isomorphism. So the map $\check{H}_2(\overline{U}/A) \to \check{H}_2(\overline{D}/S^1)$ (induced by f) is also an isomorphism. To see this, note that $\overline{U}/A = \hat{U}$ and apply the above observation. □

The rest of this section is devoted to the proof of the following theorem.

Theorem 5.14. *If* A *is either a locally separating simple arc or a separating simple closed curve in a compact* 2*-coherent space* X, *then there exists a split, degree-one map* $f : (X, A) \to (S^2, B)$ *which is bijective over* B, *where* B *is either* I *or* S^1.

The principal ingredients in the proof are Proposition 5.13 and the *strong generalized Tietze extension theorem* (SGTE) stated below.

Theorem 5.15 (SGTE). *If* A *is a closed subset of a compact metrizable space* X, *then any map* $f : A \to S^{n-1}$ $(n \in \mathbb{N})$ *has a split extension* $g : (X, A) \to (B^n, S^{n-1})$. *Furthermore, that extension is unique up to admissible homotopy.*

Proof. The so-called *generalized Tietze extension theorem* guarantees an extension $h : (X, A) \to (B^n, S^{n-1})$ (which however is not, in general, split). Define g as follows. First choose a metric ρ for X and define a second metric ρ' by $\rho'(x, y) = \min\{1, \rho(x, y)\}$. Now let $g(x) = (1 - \rho'(x, A)) \cdot h(x)$ (where B^n is viewed as

vectors of norm at most one in $\mathbb{R}^n$, and the dot in the preceding equation indicates scalar multiplication).

Now suppose that g_0 and g_1 are two such split extensions of f. Define $\varphi : (A \times [0, 1]) \cup (X \times \{0, 1\}) \to (S^{n-1} \times [0, 1]) \cup (B^n \times \{0, 1\})$ by

$$\varphi(x, t) = \begin{cases} (f(x), t) & \text{if } x \in A, \\ (g_i(x), i) & \text{if } x \in X \text{ and } i \in \{0, 1\}. \end{cases}$$

The desired homotopy is a split extension of φ to

$$(X \times I, (A \times I) \cup (X \times \{0, 1\})) \to (B^n \times I, \partial(B^n \times I)). \qquad \square$$

Proof of Theorem 5.14. We consider only the case in which A is an arc (the argument for A a closed curve is similar and easier). By Proposition 5.12 we can choose two dichotomous neighborhoods V and U of $\mathring{A}$ in X such that $\overline{U} \subset V \cup \partial A$. Hence, if U_1 and U_2 are the components of $U - A$, we have that $\operatorname{Fr} U_1 - A$ and $\operatorname{Fr} U_2 - A$ are disjoint (since they are in different components of $V - A$). Let C_1 and C_2 denote $\operatorname{Fr} U_1 - \mathring{A}$ and $\operatorname{Fr} U_2 - \mathring{A}$, respectively (note that $\operatorname{Fr} U = C_1 \cup C_2$ and $C_1 \cap C_2 = \partial A$). Let $\alpha : A \to I$ be any homeomorphism and apply the SGTE (Theorem 5.15) to $\alpha|\partial A : \partial A \to \partial I$ to obtain a split map $\beta : (C_1, \partial A) \to (S^1 \cap \{(x_1, x_2) \in \mathbb{R}^2 \subset S^2 \mid x_2 \geq 0\}, \partial I)$. Let B^2_+ and B^2_- be the upper and lower 2-disks in B^2 containing I in their boundaries. Apply the SGTE again to extend $\alpha \cup \beta : \operatorname{Fr} U_1 \to S^2$ to obtain a split map $\varphi_1 : (\overline{U}_1, \operatorname{Fr} U_1) \to (B^2_+, \partial B^2_+)$. Similarly we obtain $\varphi_2 : (\overline{U}_2, \operatorname{Fr} U_2) \to (B^2_-, \partial B^2_-)$. We have $\varphi_1|A = \varphi_2|A = \alpha$. Denote $\varphi = \varphi_1 \cup \varphi_2$. Apply SGTE a final time to extend $\varphi|\operatorname{Fr} U : \operatorname{Fr} U \to S^2$ to a split map $\psi : (\overline{X - U}, \operatorname{Fr} U) \to (S^2 - \mathring{B}^2, \partial B^2)$. Then $\varphi \cup \psi$ is the desired map.

It remains only to verify that the degree of $\varphi \cup \psi$ is one. To see this, consider the end-point compactification of the map $\varphi|U : U \to \mathring{B}^2$ which we denote by $\eta : \hat{\hat{U}} \to Q$. The target is a 2-sphere and the domain a 2-coherent space by Proposition 5.3. It follows easily from the definition of 2-coherence that η and $\varphi \cup \psi$ have the same degree. Denote by L the closure of $\mathring{I}$ in Q (so L is a simple closed curve) and note that $\eta^{-1}(L)$ is the union of a closed set of dimension zero and the closure of $\mathring{A}$ in $\hat{\hat{U}}$ (which must be either a simple arc or a simple closed curve). Since $\eta^{-1}(L)$ must separate $\hat{\hat{U}}$, we conclude from Lemma 5.6 that the closure of $\mathring{A}$ in $\hat{\hat{U}}$ is a simple closed curve. Apply the degree-one proposition to η to complete the proof. $\square$

6. A reduction of the factor theorem

In this section we show that the following result, whose proof is deferred to Section 7, implies the factor theorem. The crucial difference between the factor theorem and this factor reduction theorem is that in the former the complement of the arc (in the intermediate space) is an open Pontryagin surface whereas in the latter the analogous space is a genuine surface.

Theorem 6.1 (factor reduction theorem). *Suppose Y is a compact, connected, metrizable space, A is a closed subset of Y such that $Y - A$ is an open surface, each component of which is orientable, and $f : (Y, A) \to (S^2, C)$ is a split, surjective map which is injective over C and such that one of the following conditions is satisfied*:

(1) *$C = I$, A is 2-sided in Y, $Y - A$ has one end (hence is connected), and $f|Y - A : Y - A \to S^2 - C$ has degree one.*

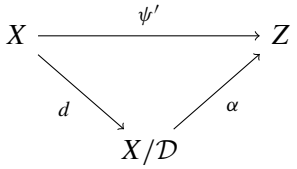
(2) *$C = S^1$, A is 2-sided in Y, and if R denotes either component of $S^2 - C$, then $f^{-1}(R)$ has one end and the map $f|f^{-1}(R) : f^{-1}(R) \to R$ has degree one.*

Then Y is a Pontryagin surface and A is $\mathbb{P}$-negligible.

(Note that in what follows we will verify the conclusion only for the first of the two conditions in the statement. The proof given the second condition is very similar, though slightly easier at certain points, and is left to the reader.)

We introduce some terminology which will be used only in this section.

Definition. Suppose $\psi : X \to Z$ is a map of compact, metrizable spaces and U is an open subset of Z such that $\psi^{-1}(U)$ is a Pontryagin surface with a sufficient family $\mathcal{E}$ which is null in X. If there exist $\mathcal{D} \subset \mathcal{E}$ with $\mathcal{D}$ sufficient for $\psi^{-1}(U)$ and a homotopy of ψ supported in U to a map ψ' having a factorization



(where d is the decomposition map), then we say α is a *Euclideanization of ψ over U using $\mathcal{E}$*. (Note that the existence of the factorization for ψ' is equivalent to the condition that for all $e \in \mathcal{D}$, $\psi'(e)$ is a singleton.)

Lemma 6.2. *Suppose $\psi : (Y, C) \to (B, \partial B)$ is a split map where B is a 2-disk, Y is a compact, metrizable space, and $\psi^{-1}(\mathring{B})$ is a Pontryagin surface with sufficient family $\mathcal{E}$ which is null in Y. Then ψ has a Euclideanization over $\mathring{B}$ using $\mathcal{E}$.*

Proof. Let $d : Y \to Y/\mathcal{E}$ be the decomposition map, and note that d is injective over $d(C)$. Apply the SGTE (Theorem 5.15) to $\psi \circ d^{-1}|d(C) : d(C) \to \partial B$ to obtain a split map $\alpha : (Y/\mathcal{E}, d(C)) \to (B, \partial B)$. To obtain the homotopy, apply the uniqueness provision of the SGTE to the maps ψ and $\alpha \circ d$ □

Proposition 6.3. *Suppose that X is a connected, compact, metrizable space, $\varphi : (X, A) \to (S^2, I)$ is a split surjective map which is one-to-one over I, $X - A$ is an open Pontryagin surface, and $\mathcal{E}$ is a sufficient family for $X - A$ which is null in X. Then φ has a Euclideanization over $S^2 - I$ using $\mathcal{E}$.*

Proof. The idea is to choose two 2-disks in S^2 such that the union of their interiors is $S^2 - I$ and then Euclideanize the map over each of the 2-disks in succession. Some care is required.

Choose three disks D, E, and F in S^2 such that $F \subset E \subset D$ and any two of the disks have boundaries intersecting in I as shown below, where I is the horizontal line segment and ∂D is the outermost simple closed curve:

From Proposition 4.1 we have that $\mathcal{E}(\varphi^{-1}(\mathring{D}))$ is sufficient for $\varphi^{-1}(\mathring{D})$. Apply Lemma 6.2 to the map

$$\varphi|(\varphi^{-1}(D), \varphi^{-1}(\partial D)) : (\varphi^{-1}(D), \varphi^{-1}(\partial D)) \to (D, \partial D)$$

to obtain a Euclideanization α of that map using $\mathcal{E}(\varphi^{-1}(\mathring{D}))$. Let $d : X \to X/\mathcal{E}(\varphi^{-1}(\mathring{D}))$ be the decomposition map and denote $\mathcal{D} = \{d(e) \mid e \in \mathcal{E} - \mathcal{E}(\varphi^{-1}(\mathring{D}))$ and $d(e) \in \alpha^{-1}(S^2 - E)\}$. By Lemma 4.2, $\mathcal{D}$ is a sufficient family for $\alpha^{-1}(S^2 - F)$ (where in the application of that lemma we use $U = \alpha^{-1}(S^2 - I)$, $A = \alpha^{-1}(F - I)$, and $B = \alpha^{-1}(E - I)$). Now Euclideanize α over $S^2 - F$ using $\mathcal{D}$. □

Proof that the factor reduction theorem implies the factor theorem. As usual we consider only the case in which the set $A \subset Q$ in the hypothesis of the factor theorem is a simple arc. Note at the outset that monotonicity of f implies monotonicity of φ. As a result, $\varphi^{-1}(A)$ must be 2-sided in X. Set $A' = \varphi^{-1}(A)$.

Let $\mathcal{E}$ be a sufficient family for P. Hence, $\mathcal{E}(P - f^{-1}(A))$ is sufficient for $P - f^{-1}(A)$. Form $\mathcal{E}' = \{\psi(e) \mid e \in \mathcal{E}(P - f^{-1}(A))\}$ and note that $\mathcal{E}'$ is sufficient for $X - A'$ (since the restriction of ψ to $P - f^{-1}(A)$ is 1-1) and is null in X. Here ψ restricts to a homeomorphism

$$P - f^{-1}(A) \to X - \varphi^{-1}(A) = X - A'.$$

Hence, both the end-point and one-point compactifications of $X - A'$ have $\check{H}_2$ isomorphic to $\mathbb{Z}$. As $f = \varphi \circ \psi$ has degree one, $\varphi : (X, A') \to (Q, A)$ must also have degree one.

Use Theorem 5.14 to obtain a split, degree-one map $\alpha : (Q, A) \to (S^2, I)$. Its composition with φ yields a degree-one map $\alpha \circ \varphi : (X, A') \to (S^2, I)$.

Apply Proposition 6.3 to obtain a Euclideanization $\varphi' : X/\mathcal{D} \to S^2$ of $\alpha \circ \varphi$ over $S^2 - I$ using $\mathcal{E}'$. Let $d : X \to X/\mathcal{D} = Y$ be the decomposition map and let $A^* = d(A')$. Note that degree φ' = degree $\alpha \circ \varphi = 1$. We leave it to the reader to verify that A^* is 2-sided in Y (hint: use the fact that the image under d of a dichotomous neighborhood of $\mathring{A}'$ in X which is saturated with respect to $\mathcal{D}$ must be a dichotomous neighborhood of A^* in Y).

Consequently, the hypotheses of the factor reduction theorem are satisfied. From it we conclude that Y is a Pontryagin surface and A^* is $\mathbb{P}$-negligible in Y. By a standard general position argument, we can choose a sufficient family $\mathcal{F}$ for Y, each element of which is disjoint from $A^* \cup \{d(e) \mid e \in \mathcal{D}\}$. It follows easily that $\{d^{-1}(e) \mid e \in \mathcal{F}\} \cup \mathcal{D}$ is a sufficient family for X, no element of which meets A'. □

7. Proof of the factor reduction theorem

We prove the factor reduction theorem (Theorem 6.1) by first producing a convergent sequence of admissible homotopies, starting with the map of the hypothesis, which progressively enlarge the 1-dimensional subspace of S^2 over which the map is bijective. The initial step (the "arc proposition") is the most difficult. It produces an admissible homotopy of the map of the hypothesis to a map which is bijective over the union of C (either I or S^1) and an arc meeting C in a preassigned point. In the proofs of both the arc proposition and the subsequent factor reduction theorem, we consider only the case $C = I$. We leave it to the reader to make the modifications necessary for the case $I = S^1$. In what follows recall that H denotes closed upper half-space in $\mathbb{R}^2$.

Proposition 7.1 (arc proposition). *Let $f : (Y, A) \to (S^2, I)$ be a map as in the hypothesis of the factor reduction theorem. Given $\varepsilon > 0$ and $a \in (-1, +1)$ there exists a homotopy of f to a map g supported in a neighborhood $U \subset H$ for which* $\operatorname{diam} U < \varepsilon$ *and* $\overline{U} \cap (\mathbb{R} \times \{0\}) = \{(a, 0)\}$, *such that g is bijective over $\{a\} \times [0, r]$ for some $r > 0$.*

Before proving the arc proposition we introduce some terminology and notation. Suppose $f : X \to Y$ is a map of spaces and $V \subset U$ are subsets of Y. The pair (U, V) is *good for f* (or merely *good* when no ambiguity can result) if $f^{-1}(V)$ is contained in a component of $f^{-1}(U)$. If Y is a metric space, then the *diameter* of the pair (U, V) is the diameter of U. For a point p in a metric space (Y, ρ), $B[p, \varepsilon]$ will denote the closed ball of ρ-radius ε centered at p.

The proof of the arc proposition requires the following five lemmas. The proofs of all but the last are left to the reader.

Lemma 7.2. *Suppose $f : X \to Y$ is a map from a compact nice space to a metric space which is one-to-one over the singleton $\{p\}$ in Y. Then given $\varepsilon > 0$ there exists $\delta > 0$ such that $(B[p, \varepsilon], B[p, \delta])$ is good.*

Lemma 7.3. *Suppose X is a nice space and $A \subset U \subset X$ with A compact and U open and connected. Then there exists a compact, connected subset C of X with $A \subset C \subset U$.*

Lemma 7.4. *Suppose $f : R \to Q$ is a proper, boundary-preserving map of surfaces and (D_1, D_2) is a pair of 2-disks in Q satisfying the following:*

(1) $D_2 \subset \mathring{D}_1 \subset D_1 \subset \mathring{Q}$.

(2) *Both* $f^{-1}(D_1)$ *and* $f^{-1}(D_2)$ *are surfaces.*

Then, if C *denotes the union of the components of* $f^{-1}(D_1)$ *meeting* $f^{-1}(D_2)$, *there exists a homotopy of* f *supported in any preassigned neighborhood of* D_1 *to a map* g *such that* $g^{-1}(D_2) = C$.

Lemma 7.5. *Any degree-one, split map* $\varphi : (Q, \partial Q) \to (B^2, \partial B^2)$ *of a compact, connected, orientable surface* Q *is admissibly homotopic to a map which is bijective over a preassigned disk in* $\mathring{B}^2$.

Note. Lemma 7.5 follows easily from the classification theorem for compact, orientable surfaces and is also an immediate consequence of the principal theorem in [Epstein 1966].

Lemma 7.6. *Suppose* Q *is a noncompact, one-ended, connected, orientable surface such that* ∂Q *has one noncompact component,* $\varphi : (Q, \partial Q) \to (H, \partial H)$ *is a proper, degree-one, split map, and* Ω *is a collar on* ∂H *in* H. *Then* φ *is properly, admissibly homotopic to a map which is bijective over* $H - \Omega$.

Proof. We will treat the case where Q has infinitely many boundary components and infinitely many handles; strategies for dealing with the other possibilities can be inferred from what we do in that slightly more complicated case. For definiteness we assume that $\Omega = \mathbb{R} \times [0, 3] \subset H$. The proof requires some care because ∂Q has two ends while Q has only one.

Let S be the subset of $\mathbb{Z}$ determined as follows (and here we denote by L the noncompact component of ∂Q). If there exists a real number b such that $\varphi(\partial Q - L) \subset (b, +\infty) \subset \partial H$, then $S = \mathbb{N}$; if there exists a number b' such that $\varphi(\partial Q - L) \subset (-\infty, b')$, then $S = \mathbb{Z} - \mathbb{N}$; otherwise, $S = \mathbb{Z}$. Now for each $n \in S$ let D_n be a 2-disk in the interior of $[n, n+1] \times [0, 1]$ and let E_n be a 2-disk in the interior of $[n, n+1] \times [2, 3]$. Also for each $n \in S$, let T_n be a punctured torus in $[n, n+1] \times [2, 3] \times [0, 1]$ with $T_n \cap H = \partial T_n = \partial E_n$ and let A_n be an annulus in $[n, n+1] \times [0, 1] \times [0, 1]$ with one component of ∂A_n equal to ∂D_n, $\partial A_n - \partial D_n \subset (n, n+1) \times \{0\} \times (0, 1)$, and $\mathring{A}_n \subset (n, n+1) \times (0, 1) \times (0, 1)$. Denote $Q' = \big[H - \bigcup_{n\in S}(\mathring{D}_n \cup \mathring{E}_n)\big] \cup \big[\bigcup_{n\in S}(A_n \cup T_n)\big]$. Let $\Phi : Q' \to H$ be the restriction to Q' of the projection map $H \times [0, 1] \to H$. Note that Φ is an admissible map.

By the classification theorem for noncompact surfaces [Prishlyak and Mischenko 2007; Richards 1963], there is a homeomorphism $\theta : Q \to Q'$. Modify θ, if necessary, so that $\Phi\theta|L, \varphi|L : L \to \partial H$ are properly homotopic. Then modify further so that, for each compact component J of ∂Q, $\Phi\theta(J) \subset [0, +\infty) \subset \partial H$ if and only if $\varphi(J) \subset [0, +\infty)$. Now it follows that the straight line homotopy μ_t is a proper homotopy between $\Phi\theta|\partial Q, \varphi|\partial Q : \partial Q \to \partial H$ (as maps to ∂H). By construction, $\Phi\theta$ is injective over $H - \Omega$.

Pass to the one-point compactifications $\hat{Q}$, $\hat{H}$, and $\partial\hat{Q}$ of Q, H, and ∂Q, respectively (the third of these is an admitted abuse of notation), and observe that $\hat{H}$ is a 2-cell. Name compactification points ∞ and ∞' in $\hat{Q}$ and $\hat{H}$, respectively. Let $A \subset \hat{Q} \times [0, 1]$ be the subset $(\hat{Q} \times \{0, 1\}) \cup (\partial\hat{Q} \times [0, 1])$. Define a map $f : A \to \hat{H} \times [0, 1]$ as $\Phi\theta$ on $\hat{Q} \times 0$, φ on $\hat{Q} \times 1$, and μ_t on $\partial Q \times [0, 1] \subset \partial\hat{Q} \times [0, 1]$, and as the map $(\infty, t) \to (\infty', t)$ on $\infty \times [0, 1]$. Apply the SGTE (Theorem 5.15) to extend f to a split map $F : (\hat{Q} \times [0, 1], A) \to (\hat{H} \times [0, 1], \partial(\hat{H} \times [0, 1]))$. A restriction of F gives a proper homotopy between $\Phi\theta$ and φ. □

Proof of the arc proposition. Note first that we can assume that ε is small enough so that $f^{-1}(B[(a, 0), \varepsilon])$ is contained in a dichotomous neighborhood of $\mathring{A}$. We will also assume without loss of generality that f is transverse to all subsurfaces of $S^2 - I$ constructed below. We denote, for $a \in (-1, +1)$ and $\delta > r > 0$, $M(a, \delta, r) = B[(a, 0), \delta] \cap \{(x, y) \in \mathbb{R}^2 \mid y \geq r\}$.

Claim 1. Given $a \in (-1, +1)$ and $\varepsilon > 0$ there exists $\delta \in (0, \varepsilon)$ such that for any $r < \delta$ there exists $s < r$ so that $(M(a, \varepsilon, s), M(a, \delta, r))$ is good.

Proof. From Lemma 7.2 there exists $\delta > 0$ such that $(B[(a, 0), \varepsilon], B[(a, 0), \delta])$ is good. Since the preimages of these sets lie in a dichotomous neighborhood of $A - \partial A$ we can conclude that $(B[(a, 0), \varepsilon] \cap \mathring{H}, B[(a, 0), \delta] \cap \mathring{H})$ is also good. Hence, $(B[(a, 0), \varepsilon] \cap \mathring{H}, M(a, \delta, r))$ is good. To finish the proof of Claim 1, apply Lemma 7.3 to conclude that, for some $s > 0$, $(M(a, \varepsilon, s), M(a, \delta, r))$ is good. □

Now continuing with the proof of the arc proposition, choose decreasing sequences $\{\delta_i\}$, $\{r_i\}$ in $(0, +\infty)$ converging to 0 and such that, for all i, $\delta_{i+1} > r_i > \delta_{i+2}$.

Denote $N_i = M(a, \delta_i, r_i)$ and observe that, for all i, $N_i \cap N_{i+1}$ is a disk, $N_i \cap N_j = \varnothing$ if $|i - j| > 1$, and $\lim(\operatorname{diam} N_i) = 0$. From Claim 1 we can find (after, in general, deleting the first K entries for some $K \in \mathbb{N}$ and reindexing) a sequence $\{(\varepsilon_i, s_i)\}_{i=1}^{\infty}$ of pairs of positive real numbers such that $\lim \varepsilon_i = 0$ and, for all i, $\varepsilon_i \geq \varepsilon_{i+1}$, $\varepsilon_i > s_i$, $r_i > s_i$, and denoting $M_i = M(a, \varepsilon_i, s_i)$, the pair (M_i, N_i) is good. Denote by Y and Z the components of $[-1, +1] \times [0, +\infty) - \overline{\bigcup_i M_i}$ containing $\{(x, 0) \mid -1 < x < a\}$ and $\{(x, 0) \mid a < x < 1\}$, respectively. Applying Claim 1 infinitely many times we can choose sequences $\{(A_i, B_i)\}$ and $\{(C_i, D_i)\}$ of good pairs of disks in $\mathring{Y}$ and $\mathring{Z}$, respectively, satisfying the following conditions:

(1) For all $i \neq j$, $A_i \cap A_j = \varnothing$ and $C_i \cap C_j = \varnothing$.

(2) Given $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $A_i \cup C_i \subset B[(a, 0), \varepsilon]$ whenever $i \geq N$.

Now by Lemmas 7.4 and 7.5, we can assume, without loss of generality, that for all i, f is bijective over $B_i \cup D_i$.

Let R be a closed subset of $H - I$ satisfying the following conditions:

(1) The closure of R in H is $R \cup \{(a, 0)\}$ and that closure is a disk.

(2) For every i, ∂R meets each of B_i and D_i transversely in an arc.

(3) $\bigcup_i M_i \subset \mathring{R}$.

Denote $Q = f^{-1}(R)$. From bijectivity of f over $\left(\bigcup_i B_i\right) \cup \left(\bigcup_i D_i\right)$ we conclude that ∂Q has one noncompact component. Denote by Q_0 that component of Q containing the noncompact component of ∂Q. Then $f|Q_0 : Q_0 \to R$ has degree one and is therefore surjective. So for all i, $f(Q_0) \cap N_i \neq \varnothing$ and hence, from goodness of the pairs $\{(M_i, N_i)\}$ we have $f(Q - Q_0) \cap \left(\bigcup_{i=1}^{\infty} N_i\right) = \varnothing$. Now, because $R - \bigcup_{i=1}^{\infty} N_i$ is homeomorphic to $\mathbb{R} \times [0, 1)$, all components of $Q - Q_0$ can be "eliminated" (i.e., in the image, "pushed out of" R) by an admissible homotopy of f fixing f outside any preassigned neighborhood of $Q - Q_0$ in $X - A$ (the details of this argument are left to the reader). So we have established the following claim.

Claim 2. We can assume without loss of generality that Q is connected.

We will show that we can also assume without loss of generality that Q has one end which, by Lemma 7.6, will complete the proof of the arc proposition. To establish this, let $\{W_n\}_{n=1}^{\infty}$ be an exhaustion of R satisfying the following conditions, where Z_n denotes the closure in R of $R - W_n$:

(1) For all n, W_n is a disk such that $W_n \cap \partial R$ is an arc and $W_n \subset \operatorname{Int} W_{n+1}$ (where the interior is with respect to R).

(2) Given n there exists j such that

$$\partial W_n \cap \left(\bigcup_{i=1}^{\infty} N_i\right) = \partial W_n \cap [N_j - (N_{j+1} \cup N_{j-1})]$$

and this set is an arc.

(3) If, for some n and i, $Z_n \cap N_i \neq \varnothing$, then $M_i \subset Z_{n-1}$.

Now, by an argument similar to that which established Claim 2, we can assume the following without loss of generality: $(*)$ for all n, no component of $f^{-1}(Z_n)$ is sent by f into $Z_n - \left(\bigcup_{i=1}^{\infty} N_i\right)$ (we leave this to the reader, but note first that the closure of $Z_n - \left(\bigcup_{i=1}^{\infty} N_i\right)$ in R has two components, each of which is homeomorphic to H).

Now, for a given n, there must exist a component C of $f^{-1}(Z_n)$ such that $f|C : C \to Z_n$ has nonzero degree and is therefore surjective. If C' is another component of $f^{-1}(Z_n)$, we have by $(*)$ that, for some $j \in \mathbb{N}$, $f(C') \cap N_j \neq \varnothing$. By condition (3) for $\{W_n\}$ we then have $M_j \subset Z_{n-1}$ and hence, by goodness of the pair (M_j, N_j), we have that $C' \cup C$ is contained in a component of $f^{-1}(Z_{n-1})$. Hence, Q has one end. □

Notation. To avoid ambiguity in the sequel, the notation (a, b) (where $a, b \in \mathbb{R}$) will be used exclusively for open intervals in $\mathbb{R}$. The map $p_2 : \mathbb{R}^2 \to \mathbb{R}$ is projection to the second coordinate.

In the observations below (used in the proof of Lemma 7.7), R is a compact, orientable surface with boundary.

Observation. (1) *If two split maps from* $(R, \partial R)$ *to* $(B^2, \partial B^2)$ *are equal over* ∂B^2, *then they are admissibly homotopic.*

(2) *If* $0 < a < b < 1$ *and* $\varphi : \partial R \to \partial([0,1]^2)$ *is a map bijective over* $\{0,1\} \times [0,1]$, *then* φ *extends to a split map from* $(R, \partial R)$ *to* $([0,1]^2, \partial([0,1]^2))$ *which is bijective over* $[0,1] \times [a,b]$.

Proof. (1) The straight line homotopy between f and g is admissible.

(2) We leave this to the reader except noting that we can assume without loss of generality that R is planar. To see this, first use the classification of compact surfaces to show that $\mathring{R}$ contains a compact surface S such that R/S is a planar surface. □

Lemma 7.7. *Suppose* R *is a connected, orientable, noncompact surface having one end. Also suppose that* $\varphi : (R, \partial R) \to (Q, \partial Q)$ *is a proper, split map where* $Q = [a,b] \times (0,c]$ (*for some* a, b, c *with* $a < b$ *and* $c > 0$) *which is bijective over* $\{a,b\} \times (0,c]$. *Then there exists* $s < c$ *such that for any* $t < s$ *and* $0 < \varepsilon < \frac{c-t}{2}$ *there is a proper, admissible homotopy of* φ *supported in* $(a,b) \times (t,c)$ *to a map which is bijective over* $[a,b] \times [t+\varepsilon, c-\varepsilon]$.

Addendum. There is a straightforward generalization of Lemma 7.7 which we will need. In the hypothesis of that generalization connectivity of R is replaced by the following: R has finitely many components only one of which is noncompact. Then in the conclusion the number c is replaced by r with $0 < r < c$ such that no compact component of R meets $\varphi^{-1}([a,b] \times (0,r])$.

We leave the full statement and proof of the generalization to the reader. In the sequel, it will be understood that "Lemma 7.7 " refers to this generalization.

Proof. By one-endedness of R we can choose s so that if $t \leq s$, then the image of only one component of $\varphi^{-1}([a,b] \times [t,c])$ meets both components of $[a,b] \times \{t,c\}$. Let C denote that component. We can assume without loss of generality that the images of all other components of $\varphi^{-1}([a,b] \times [t,c])$ are in $[a,b] \times ([t, t+\varepsilon) \cup (c-\varepsilon, c])$. We leave it to the reader to complete the proof using the above observations. □

Now for the remainder of the proof of the factor reduction theorem we adopt the following notation: for $a > 0$ and $n \in \mathbb{N}$, $E_n = \left\{ \frac{k}{2^n} \;\middle|\; k \in \mathbb{Z} \text{ and } \left|\frac{k}{2^n}\right| < 1 \right\}$, $M_n = \max E_n$, $W\langle n, a\rangle = [-M_n, M_n] \times (0,a]$, and $Z\langle n, a\rangle = (E_n \times (0,a]) \cup ([-M_n, M_n] \times \{a\})$; for $b < 0$, $Z\langle n, b\rangle = \{-\vec{x} \mid \vec{x} \in Z\langle n, -b\rangle\}$ (and similarly for $W\langle n, b\rangle$).

We will show that, for some strictly decreasing sequence $\{a_n\}_{n=1}^{\infty}$ converging to zero, f is admissibly homotopic to a map which is bijective over $\bigcup_{k=1}^{\infty} Z\langle k, a_k\rangle$. We construct such a map as the limit of a sequence $\{f_n\}_{n=1}^{\infty}$ of split maps where, for

each n, f_n is bijective over $\bigcup_{k=1}^{n} Z\langle k, a_k\rangle$ (for appropriately chosen $\{a_1, a_2, \ldots, a_n\}$) and f_{n+1} is admissibly homotopic to f_n. The following observation will be used to ensure convergence of $\{f_n\}_{n=1}^{\infty}$.

Observation. *A sequence $\{g_n : (Y, A) \to (S^2, I)\}_{n=1}^{\infty}$ of admissible maps converges to a map admissibly homotopic to g_1 if, for each $n > 1$, g_{n-1} is homotopic to g_n by a homotopy supported in an open subset U_n of $S^2 - I$ having finitely many components and compact closure in $S^2 - I$ and satisfying the following: for all $n \neq m$, $U_n \cap U_m = \varnothing$, and*

$$\lim_{n\to\infty} \max\{\operatorname{diam}(\bar{C}) \mid C \text{ is a component of } U_n\} = 0.$$

Now, applying the arc proposition infinitely many times, construct a monotone strictly decreasing sequence $\{b_n\}_{n=1}^{\infty}$ of real numbers converging to zero and a map f_0 admissibly homotopic to f and bijective over $\bigcup_{n=1}^{\infty} E_n \times [0, b_n]$. The map f_0 is itself the limit of a sequence of maps each of which is obtained by an application of the arc proposition to its predecessor. The above observation is used to ensure convergence and to verify that the limit is admissibly homotopic to f. The details are left to the reader.

Now given f_0 we will construct f_1, then briefly indicate the construction of f_2. The induction step in full generality we leave to the reader.

To construct f_1, let F denote the closure in $H - \partial H$ of one of the components of $W\langle 1, b_1\rangle - Z\langle 1, b_1\rangle$ (which is $\left[\left(-\frac{1}{2}, 0\right) \cup \left(0, \frac{1}{2}\right)\right] \times (0, b_1)$). Apply Lemma 7.7 to $f_0 | f_0^{-1}(F) : f_0^{-1}(F) \to F$ for each choice of F. Choose the r and t (as in Lemma 7.7) to be the same for both applications. We are free to choose r small enough so that each of the two homotopies has support in an open rectangle whose diameter is less than one. Let f_1 be the map which results from the composition of the two homotopies. From the conclusion of Lemma 7.7 we can choose $a_1 < b_1$ such that f_1 is bijective over $\left[-\frac{1}{2}, \frac{1}{2}\right] \times \{a_1\}$.

Now to construct f_2, apply Lemma 7.7 to each map $f_1 | f_1^{-1}(F) : f_1^{-1}(F) \to F$ where $F = \left[\frac{k}{2^2}, \frac{k+1}{2^2}\right] \times (0, c_2]$ where $c_2 = \min\{a_1, b_2\}$ and k is an integer with $-3 \leq k \leq 2$. Choose a common r and t for the six applications of the lemma and furthermore choose r small enough so that each of the homotopies has support in a rectangle of diameter one half and furthermore that support is disjoint from the support of the previously constructed homotopy of f_0 to f_1. The composition of the six homotopies is the homotopy from f_1 to f_2. The conclusion of Lemma 7.7 allows us to choose $a_2 < c_1$ such that f_2 is bijective over $Z\langle 1, a_1\rangle \cup Z\langle 2, a_2\rangle$.

By "symmetry" we can now assume without loss of generality that the map f of the hypothesis is bijective over $\bigcup_{n=1}^{\infty}[Z\langle n, a_n\rangle \cup Z\langle n, b_n\rangle]$ where $\{a_n\}$ and $\{b_n\}$ are monotone strictly decreasing and monotone strictly increasing, respectively, and both sequences converge to zero. Note that the closure in S^2 of $\bigcup_{n=1}^{\infty}[W\langle n, a_n\rangle \cup W\langle n, b_n\rangle]$ is a 2-disk and denote the closed complement of that 2-disk minus the

endpoints of I by Q (it is homeomorphic to $[-1,+1]\times\mathbb{R}$). Denote $R=f^{-1}(Q)$. We have that $f|R: R\to Q$ is a proper map of noncompact surfaces and $f|\partial R:\partial R\to\partial Q$ is bijective. Furthermore, it follows from one-endedness of $Y-A$ that $f|R: R\to Q$ is bijective on ends. It follows from the classification theorem for noncompact surfaces [Richards 1963; Prishlyak and Mischenko 2007] (and the special case required here can also be proven by applying Lemma 7.7 infinitely many times) that R can be constructed by first deleting the interiors of a proper family of pairwise-disjoint 2-disks in $\mathbb{R}\times[0,1]$, none of which meets $\mathbb{R}\times\{0,1\}$, and then attaching to the boundary of each 2-disk a once-punctured torus. Denote by S the decomposition space obtained from the decomposition of R whose only nondegenerate elements are the punctured tori. Up to admissible homotopy fixing $f|\partial R$, the map $f|R: R\to Q$ factors through a map $g: S\to Q$ whose end-point compactification is a split boundary-to-boundary map of 2-disks. Hence, from the observation preceding Lemma 7.7, that map is admissibly homotopic (fixing $g(\partial S)$ to a homeomorphism). So we can assume without loss that the map f sends each punctured torus to a point and is injective over the complement of the image of the union of all the punctured tori. So we can easily choose a proper, split embedding $(\mathbb{Z}\times[-1,+1],\mathbb{Z}\times\{-1,+1\})\to(Q,\partial Q)$ such that f is injective over the image of the embedding (which we denote by E). So now we can assume without loss of generality that the map f is bijective over $I\cup E\cup\left(\bigcup_{n=1}^{\infty}[Z\langle n,a_n\rangle\cup Z\langle n,b_n\rangle]\right)$, which we denote by Z.

Note that the closure of any component C of the complement of Z in S^2 is a 2-disk and that $f|f^{-1}(C): f^{-1}(C)\to C$ is a boundary-preserving map of compact, connected orientable surfaces which is bijective over ∂C and hence is homotopic (fixing $f|\partial f^{-1}(C)$) to a standard map. It follows easily that Y is a Pontryagin surface and A is $\mathbb{P}$-negligible.

8. Pontryagin disks

Recall that a *Pontryagin disk* $\mathbb{D}$ is a compact, connected subset of a rich Pontryagin surface P whose frontier relative to P is a simple closed curve. That curve is called the *boundary of* $\mathbb{D}$ and is denoted $\partial\mathbb{D}$. The subset $\mathbb{D}-\partial\mathbb{D}$ is the *interior of* $\mathbb{D}$, written $\operatorname{Int}\mathbb{D}$. By Corollary 3.2 every Pontryagin disk $\mathbb{D}$ has a rich family $\mathcal{E}$ of figure-eights, all of which lie in $\operatorname{Int}\mathbb{D}$; we shall assume that every sufficient family for a Pontryagin disk used here has this property.

Theorem 8.1. *Suppose $\mathbb{D}$ and $\mathbb{D}'$ are Pontryagin disks equipped with sufficient families $\mathcal{E}$ and $\mathcal{E}'$, respectively. Let $S=\mathbb{D}/\mathcal{E}$ and $S'=\mathbb{D}'/\mathcal{E}'$ be the associated decompositions and let $d:\mathbb{D}\to S$ and $d':\mathbb{D}'\to S'$ be the quotient maps. Let Z be a closed subset of S such that $Z\cap d(\mathcal{E})=\varnothing$, and let $h: S\to S'$ be a homeomorphism such that $h(Z)\cap d'(\mathcal{E}')=\varnothing$. Then for any $\varepsilon>0$ there exists a homeomorphism $H:\mathbb{D}\to\mathbb{D}'$ such that hd and $d'H$ are ε-close and equal on $d^{-1}(Z)$.*

The proof of the above theorem, which occupies most of this section, is deferred.

Corollary 8.2. *Every homeomorphism $\psi : \partial\mathbb{D} \to \partial\mathbb{D}'$ between the boundaries of Pontryagin disks $\mathbb{D}, \mathbb{D}'$ extends to a homeomorphism $\Psi : \mathbb{D} \to \mathbb{D}'$.*

Proof. Let $\mathcal{E}$ and $\mathcal{E}'$ be full families for $\mathbb{D}$ and $\mathbb{D}'$, respectively. Recall that by convention no elements of $\mathcal{E}$ or $\mathcal{E}'$ meet $\partial\mathbb{D}$ or $\partial\mathbb{D}'$. Let $B = \mathbb{D}/\mathcal{E}$ and $B' = \mathbb{D}'/\mathcal{E}'$ denote the usual decompositions and $d : \mathbb{D} \to B$ and $d' : \mathbb{D}' \to B'$ the quotient maps. Since B and B' are 2-disks the homeomorphism $d' \circ \psi \circ (d|\partial B)^{-1} : \partial B \to \partial B'$ extends to a homeomorphism $h : B \to B'$. Apply Theorem 8.1 with $Z = \partial S$. □

Corollary 8.3. *Let J and J^* denote separating simple closed curves in closed, rich Pontryagin surfaces P and P^*, respectively. Then any homeomorphism $h : J \to J^*$ can be extended to a homeomorphism $H : P \to P^*$.*

Proof. Each of J and J^* bounds two Pontryagin disks in their respective Pontryagin surfaces. Apply Corollary 8.2. □

Theorem 8.1 also supplies an affirmative answer to a question raised by D. Repovš on several occasions back in the 1990s. The argument for Corollary 8.4 below also yields that Cantor sets in connected rich Pontryagin surfaces are homogeneously embedded.

Corollary 8.4. *Suppose that $\mathbb{D}$ and $\mathbb{D}'$ are Pontryagin disks and that $K \subset \operatorname{Int}\mathbb{D}$ and $K' \subset \operatorname{Int}\mathbb{D}'$ are Cantor sets. Then each homeomorphism $h : \partial\mathbb{D} \cup K \to \partial\mathbb{D}' \cup K'$ extends to a homeomorphism $H : \mathbb{D} \to \mathbb{D}'$.*

Proof. Here K is $\mathbb{P}$-negligible in $\operatorname{Int}\mathbb{D}$, so there exists a full collection $\mathcal{E}$ of figure-eights for $\mathbb{D}$, all of which lie in $\operatorname{Int}\mathbb{D} - K$. Similarly, there exists a full collection $\mathcal{E}'$ of figure-eights for $\mathbb{D}'$, all of which lie in $\operatorname{Int}\mathbb{D}' - K'$. Apply Theorem 8.1 using the obvious decompositions. □

Definition. Let $\mathbb{D}$, $\mathcal{E}$, and d be as in Theorem 8.1. A *utilitarian web W for S* is a finite collection $\{B_i\}$ of 2-cells in S that cover S, whose boundaries miss $d(\mathcal{E})$, and for $i \neq j$, $B_i \cap B_j$ is either empty or a connected subset of the boundary of each. (A utilitarian web is a generalized triangulation.) We define utilitarian webs W on appropriate quotients of closed Pontryagin surfaces similarly. We will refer to the union of the boundaries as the 1-*skeleton* of W. We will call two such webs W, W' for S *equivalent* if there exists a homeomorphism $h : S \to S$ that induces a bijection from the cells of W to the cells of W'.

The following is an immediate consequence of Theorem 8.1:

Corollary 8.5. *Under the hypotheses of Theorem 8.1, let W be a utilitarian web for S and let $h : S \to S'$ be a homeomorphism that carries the 1-skeleton T of W into $S' - d'(\mathcal{E}')$. Then there exists a homeomorphism $H : \mathbb{D} \to \mathbb{D}'$ such that $hd(t) = d'H(t)$ for all $t \in d^{-1}(T) \cup \partial\mathbb{D}$.*

We state several definitions before returning to the proof of Theorem 8.1.

Definition. A Pre$\mathbb{P}$ *space* X is a space equipped with two subspaces denoted $E(X)$ and $C(X)$ satisfying the following: X is a compact, connected, orientable surface with connected boundary; $E(X)$ is the union of a finite family of pairwise-disjoint figure-eights in $\mathring{X}$ such that, for each figure-eight $e \in E(X)$, X/e is a surface; $C(X)$ is a countable dense subspace of $\mathring{X}$ disjoint from $E(X)$.

A map $f : X \to Y$ of Pre$\mathbb{P}$ spaces is a Pre$\mathbb{P}$ *map* if it is standard and satisfies $\{y \in Y \mid |f^{-1}(y)| \neq 1\} \subseteq C(Y)$ and $f(E(X) \cup C(X)) = E(Y) \cup C(Y)$. Note that compositions of Pre$\mathbb{P}$ maps are Pre$\mathbb{P}$.

A *diagram* is a set $\mathcal{D}$ of surjective maps of compact metric spaces satisfying the following conditions: the range of no element of $\mathcal{D}$ is the same space as its domain; no two elements of $\mathcal{D}$ have both the same domain and the same range. A *derived* map of a diagram $\mathcal{D}$ is a map which is the composition of elements of $\mathcal{D}$ such that no element of $\mathcal{D}$ appears more than once in the factorization and the domain and range of the composite map are different (subsequently, when we refer to a "factorization" of a derived map, it will be understood that the factorization satisfies this condition).

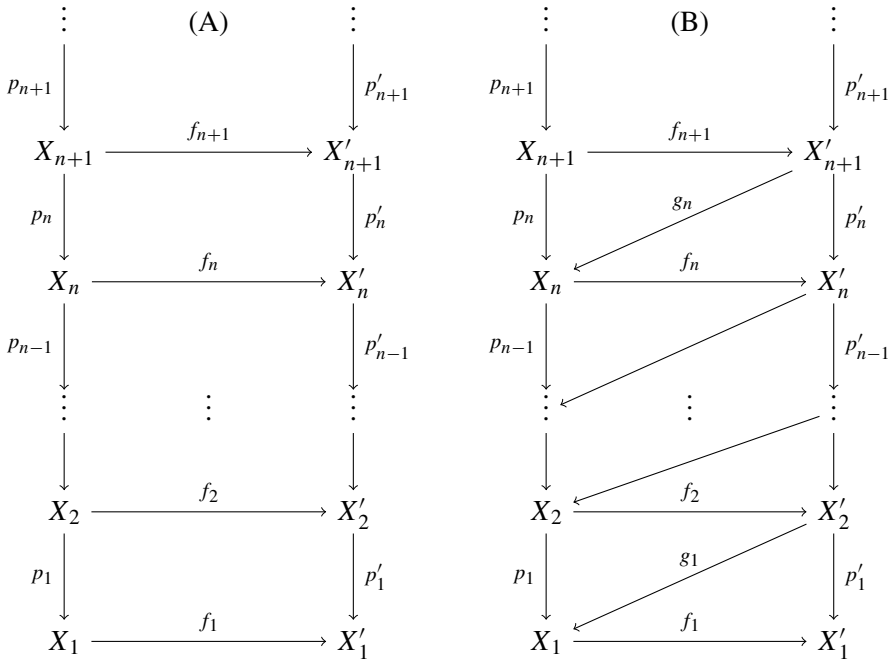
A *modulus of continuity* of a diagram is a function $\delta : (0, +\infty) \to (0, +\infty)$ which is a modulus of continuity for every derived map of the diagram (i.e., given a derived map f of the diagram, $\varepsilon > 0$, and points x and y in the domain of f which are $\delta(\varepsilon)$-close we have that $f(x)$ and $f(y)$ are ε-close). Note that any finite diagram has a modulus of continuity.

A pair $X \overset{f}{\underset{g}{\rightrightarrows}} Y$ of derived maps of a diagram is *allowable* if f and g have factorizations such that if $\mathcal{A}$ and $\mathcal{B}$ denote the sets of spaces appearing in the factorizations of f and g, respectively, then $\mathcal{A} \cap \mathcal{B} = \{X, Y\}$.

A diagram $\mathcal{D}$ is ε-*commutative* if the two maps of any allowable pair are ε-close. An infinite diagram is *asymptotically commutative* if, given $\varepsilon > 0$, there exists a finite subset $\mathcal{D}_\mathcal{E}$ of $\mathcal{D}$ such that $\mathcal{D} - \mathcal{D}_\mathcal{E}$ is ε-commutative.

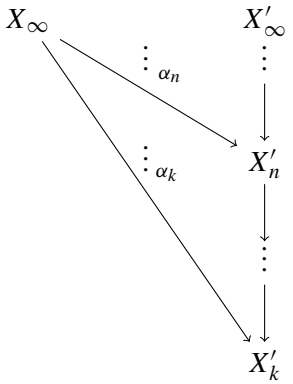
Observation. *Given an inverse sequence* $\{p_n : X_{n+1} \to X_n\}_{n=1}^{\infty}$ *of surjective maps of compact metrizable spaces with limit* X_∞, *there exist metrics* $\{\rho_n\}_{n=1}^{\infty}$ *for* $\{X_n\}_{n=1}^{\infty}$ *and* ρ_∞ *for* X_∞ *so that, for any* $x, y \in X_\infty$, *the sequence* $\{\rho_n(p_{\infty,n}(x), p_{\infty,n}(y))\}_{n=1}^{\infty}$ (*where* $p_{\infty,n} : X_\infty \to X_n$ *is the projection map*) *is strictly increasing and has limit* $\rho_\infty(x, y)$. (We leave the proof to the reader.)

Definition. Diagrams of the following two forms will be referred to as *type A* and *type B* diagrams if, in each case, the metrics in the vertical columns (which are inverse sequences) satisfy the condition stated in the above observation:



Lemma 8.6. *Given an asymptotically commutative type A diagram $\mathcal{D}$ (with notation as in the definition) there exists a map $f_\infty : X_\infty \to X'_\infty$ having the following property: for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that if $n \geq N$, then $p'_{\infty,n} \circ f_\infty$ and $f_n \circ p_{\infty,n}$ are ε-close.*

Proof. From asymptotic commutativity of $\mathcal{D}$ we have that, for any $k \in \mathbb{N}$, the sequence of maps $\{p'_{n,k} \circ f_n \circ p_{\infty,n} : X_\infty \to X'_k\}_{n>k}$ converges uniformly. Denoting the limit by α_k, we have that the diagram

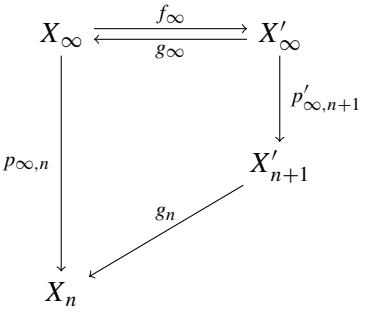


commutes. The inverse limit is f_∞. □

Note. We will refer to the map f_∞ in the conclusion of Lemma 8.6 as the *limit* of $\mathcal{D}$.

Lemma 8.7. *Suppose $\mathcal{D}$ is an asymptotically commutative type B diagram and let f_∞ and g_∞ be the limits, respectively, of the two type A diagrams obtained when the $\{g_i\}$ are deleted from $\mathcal{D}$ and when the $\{f_i\}$ are deleted from $\mathcal{D}$. Then f_∞ and g_∞ are inverses.*

Proof. To show that g_∞ is a left inverse for f_∞ it suffices to show, given $\varepsilon > 0$ and $a \in X_\infty$, that $(g_\infty \circ f_\infty)(a)$ is within ε of a. To see this, extract the subdiagram



We can choose n large enough so that the following conditions are satisfied:

(1) There exists $\delta > 0$ such that if $E \subset X_n$ has diameter less than 2δ, then $p_{\infty,n}^{-1}(E)$ has diameter less than ε.

(2) The subdiagram is δ-commutative.

By condition (2), then $(p_{\infty,n} \circ g_\infty \circ f_\infty)(a)$ is within δ of $(g_n \circ p'_{\infty,n+1} \circ f_\infty)(a)$ and the latter is within δ of $p_{\infty,n}(a)$. The desired inequality follows from condition (1).

The proof that f_∞ is a left inverse for g_∞ is similar. □

Lemma 8.8. *Suppose X is a compact metric space, R is an open surface which is an open subset of X, A and B are countable, dense subsets of R, and $\varepsilon > 0$. Then there exists a split homeomorphism $\varphi : (X, R) \to (X, R)$ supported in R and ε-close to the identity such that $\varphi(A) = B$.*

Proof. Brouwer [1913] and Fréchet [1910], independently, proved that Euclidean space is countable dense homogeneous. This is a mild generalization of their result. We provide some details for completeness.

The idea is to produce φ as a limit of a sequence $\varphi_k : X \to X$ supported in R. For each $k \geq 1$ we will determine a homeomorphism $h_k : X \to X$ supported in a very small 2-disk Δ_k in R and then will set $\varphi_k = h_k \varphi_{k-1}$ (here $\varphi_0 =$ Identity).

List the elements $\{a_1, a_2, \ldots\}$ of A and likewise the elements $\{b_1, b_2, \ldots\}$ of B.

When $k = 2m - 1$, Δ_k will be centered at $\varphi_{k-1}(a_m)$ and will contain no other point of

$$\varphi_{k-1}(\{a_1, a_2, \ldots, a_{m-1}\}) \cup \{b_1, b_2, \ldots, b_{m-1}\} \subset B.$$

If $\varphi_{k-1}(a_m) \in B$. then h_k will be the identity; otherwise, apply density of B to obtain $b_j \in B \cap \operatorname{Int} \Delta_k$ and choose h_k so $h_k \varphi_{k-1}(a_m) = b_j$.

When $k = 2m$, Δ_k will be centered at $\varphi_{k-1}(b_m)$ and contain no other point of

$$\varphi_{k-1}(\{a_1, a_2, \ldots, a_m\}) \cup \{b_1, b_2, \ldots, b_{m-1}\} \subset B.$$

If there exists $a_j \in A$ such that $\varphi_{k-1}(a_j) = b_m$, then h_k will be the identity; otherwise, apply density of A to obtain $a_j \in A \cap \operatorname{Int} \Delta_k$ and choose h_k so $h_k \varphi_{k-1}(a_j) = b_m$.

In short, at odd-numbered stages of the process, a point of A is shifted into B, in orderly fashion, and at even-numbered stages a point of B is caused to be the image of some point of A. Once such arrangements are made, no further adjustment of those special points is allowed at later stages, so those arrangements persist to the limit map φ. Eventually all points of A are moved into B and all from B are covered.

Simply by choosing the Δ_k of diameter less than $\varepsilon/2^k$, we can assure that the sequence $\{\varphi_k\}$ converges uniformly to a continuous function φ ε-close to the identity. Furthermore, φ will restrict to the identity on $X - R$ and will be surjective over X.

At any stage $k > 1$ in this process, we can determine $\eta_{k-1} > 0$ such that points of X at least $1/k$ apart have image under φ_{k-1} at least η_{k-1} apart. Thus, by requiring Δ_k to have diameter less than $\eta_i/2^{k-i}$ for $i = 1, 2, \ldots, k-1$, we assure injectivity of φ. As a result, φ is a split homeomorphism of (X, R) to itself. □

Lemma 8.9. *Suppose $f : X \to X'$ and $p : Y \to X'$ are* Pre$\mathbb{P}$ *maps such that $\{x \in X' \mid |f^{-1}(x)| \neq 1\} \subseteq \{x \in X' \mid |p^{-1}(x)| \neq 1\}$ and Z is a closed subset of X' disjoint from $E(X') \cup C(X')$. Then given $\varepsilon' > 0$ there exists a* Pre$\mathbb{P}$ *map $g : Y \to X$ such that p and $f \circ g$ are ε'-close and equal over Z.*

Proof. Denote $\{x_1, x_2, \ldots, x_n\} = \{x \in X' \mid |p^{-1}(x)| \neq 1\}$. Let $\{D_i\}_{i=1}^n$ be a pairwise-disjoint family of 2-disks in $\mathring{X}'$ such that for each i

$$x_i \in \mathring{D}_i, \qquad \operatorname{diam} D_i < \varepsilon', \qquad \partial D_i \cap [E(X') \cup C(X') \cup Z] = \varnothing.$$

Define g as follows. For $x \notin \bigcup_{i=1}^n p^{-1}(\mathring{D}_i)$ we define $g(x) = f^{-1}(p(x))$ (this can be done since f is injective over the complement of $\bigcup_i \mathring{D}_i$). For each i we define $\alpha_i = g|p^{-1}(D_i) : p^{-1}(D_i) \to f^{-1}(D_i)$ as follows. Note first that, for all i, $p^{-1}(D_i)$ is a disk with a handle and $f^{-1}(D_i)$ is either a disk with a handle or simply a disk. In the first case choose a homeomorphism α_i satisfying the following conditions:

(1) $\alpha_i|p^{-1}(\partial D_i) = f^{-1} \circ p|p^{-1}(\partial D_i)$.

(2) α_i carries $p^{-1}(x_i)$ onto $f^{-1}(x_i)$.

(3) $\alpha_i(C(Y) \cap p^{-1}(D_i)) = C(X) \cap f^{-1}(D_i)$.

Note that (2) can be achieved since the figure-eight in a disk with a handle is unique up to homeomorphism fixing the boundary and (3) can be achieved using Lemma 8.8.

In the second case ($f^{-1}(D_i)$ is a disk), just choose α_i so that $\alpha_i^{-1}(f^{-1}(x_i)) = p^{-1}(x_i)$ and $p^{-1}(x_i)$ is the only nontrivial point preimage. Again Lemma 8.8 allows us to achieve condition (3) above. □

Lemma 8.10. *Suppose $\mathcal{D}$ is a finite ε-commutative diagram, δ is a modulus of continuity for $\mathcal{D}$, and $f : Z \to Y$ and $p : X \to Y$ are maps in $\mathcal{D}$ such that X is neither the domain nor codomain of any map in $\mathcal{D}$ other than p. If $r > 0$, $\varepsilon' \leq \delta(r)$, and $g : X \to Z$ is a map not in $\mathcal{D}$ and such that $f \circ g$ is ε'-close to p, then the diagram $\mathcal{D} \cup \{g\}$ is σ-commutative where $\sigma = \max\{\varepsilon', \varepsilon + r\}$.*

Proof. Suppose $\{\varphi, \psi\}$ is an allowable pair in $\mathcal{D} \cup \{g\}$. If each of p, f, and g is a factor of φ or a factor of ψ, then they must be the only factors in the two factorizations and we are done by hypothesis. If neither of the factorizations of φ and ψ include g, then $\{\varphi, \psi\}$ is an allowable pair in $\mathcal{D}$ and again we are done. The remaining possibility is that g and p are the initial factors of φ and ψ and f is a factor of neither. We will show that, in this case, φ and ψ are $(\varepsilon + r)$-close. We can write $\varphi = \alpha \circ g$ and $\psi = \beta \circ p$ where α and β are derived maps of $\mathcal{D}$. Consider the three maps $\beta \circ p$, $\beta \circ f \circ g$, and $\alpha \circ g$. The first and second are r-close because p and $f \circ g$ are ε'-close (and $\varepsilon' \leq \delta(r)$). The second and third are ε-close since the distance between them is the same as the distance between $\beta \circ f$ and α (an allowable pair in $\mathcal{D}$). The triangle inequality concludes the argument. □

Proof of Theorem 8.1. First note that we can assume (without loss of generality) $\partial S \subset Z$. Using Proposition 2.1 we write

$$\mathbb{D} = X_\infty = \varprojlim\{p_n : X_{n+1} \to X_n\}_{n=1}^\infty,$$
$$\mathbb{D}' = X'_\infty = \varprojlim\{p'_n : X'_{n+1} \to X'_n\}_{n=1}^\infty,$$

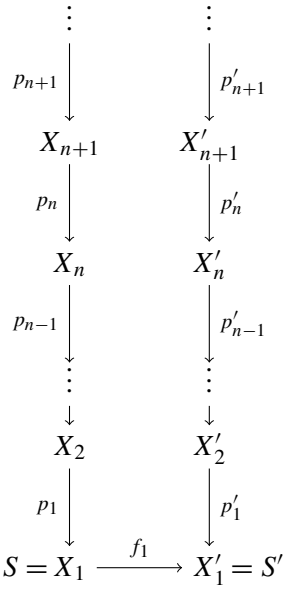
where $X_1 = S$ and $X'_1 = S'$. Furthermore, we can assume that the metrics for these spaces satisfy the conditions stated in the observation following the definition of ε-commutative. For each $n \in \mathbb{N}$ we set

$$C(X_n) = p_{\infty,n}(\{e \in \mathcal{E} \mid |p_{\infty,n}(e)| = 1\}),$$
$$E(X_n) = p_{\infty,n}(\{e \in \mathcal{E} \mid |p_{\infty,n}(e)| \neq 1\})$$

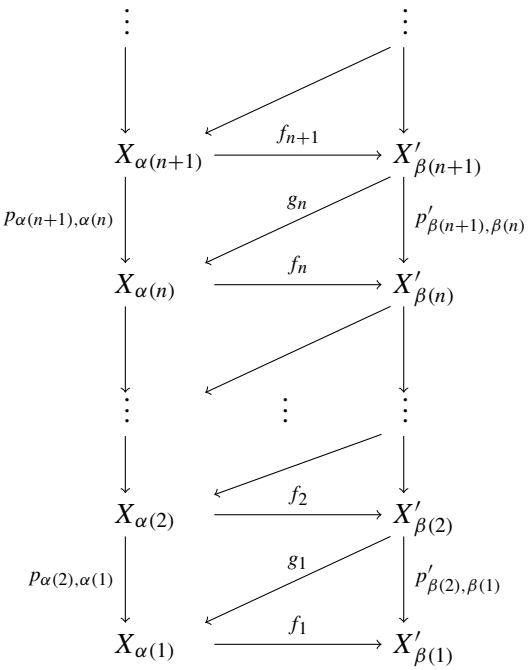
and similarly for $C(X'_n)$ and $E(X'_n)$. This makes the maps $\{p_n\}$ and $\{p'_n\}$ Pre$\mathbb{P}$. Note also that $d = p_{\infty,1}$ and $d' = p'_{\infty,1}$.

The following argument shows that h can be approximated by Pre$\mathbb{P}$ maps and hence we can assume without loss that it is Pre$\mathbb{P}$. Applying Lemma 8.8 to the pair $(X'_1, X'_1 - h(Z))$ and the subsets $h(C(X_1))$ and $C(X'_1)$ we obtain a homeomorphism $\varphi : X'_1 \to X'_1$ fixing $h(Z)$ and throwing $h(C(X_1))$ onto $C(X'_1)$ and ε-close to the identity for any preassigned ε. So $\varphi \circ h$ is Pre$\mathbb{P}$ and arbitrarily close to h. We denote

$h = f_1$ and assume f_1 is Pre$\mathbb{P}$. So we have an infinite diagram of Pre$\mathbb{P}$ maps:



We will construct the infinite diagram of Pre$\mathbb{P}$ maps (denoted $\mathcal{D}$),



where α and β are increasing functions from $\mathbb{N}$ to $\mathbb{N}$ with $\alpha(1) = \beta(1) = 1$.

Before listing the properties which $\mathcal{D}$ will have, we adopt some notation: $\mathcal{D}_k^{(n)}$ (for $1 \leq k < n$) will denote the finite subset of $\mathcal{D}$ consisting of all maps, each of which is either f_k, f_n, or a map in $\mathcal{D}$ above f_k and below f_n; define Z_n inductively by $Z_1 = Z$ and $Z_{n+1} = p_{\alpha(n+1),\alpha(n)}^{-1}(Z_n)$.

The diagram $\mathcal{D}$ will satisfy the following two conditions:

(1) For all n and $1 \leq k < n$, $\mathcal{D}_k^{(n)}$ is $\varepsilon \cdot \left[\sum_{m=k}^{n} \frac{1}{2^m}\right]$-commutative.

(2) f_n and g_n are injective over $f_n(Z_n)$ and Z_n, respectively.

It follows immediately from the above properties that $\mathcal{D}$ is asymptotically commutative and that the map f_∞ provided by Lemma 8.6 and guaranteed to be a homeomorphism by Lemma 8.7 will serve as the desired H.

The construction of $\mathcal{D}$ is accomplished by producing inductively the sequence $\{\mathcal{D}_1^{(n)}\}_{n=2}^{\infty}$ of subdiagrams. The construction of $\mathcal{D}_1^{(n+1)}$ from $\mathcal{D}_1^{(n)}$ is carried out in two stages. First $\beta(n+1)$ and g_n are chosen, then $\alpha(n+1)$ and f_{n+1}. In each stage, Lemma 8.9 is used to construct the desired map. The choices of ε' (as in Lemma 8.9) which will ensure the necessary approximate commutativity of $\mathcal{D}_k^{(n+1)} (k < n)$ are dictated by Lemma 8.10.

So in determining $\beta(n+1)$ first note that given any finite subset S of $C(X'_{\beta(n)})$ there exists $m \in \mathbb{N}$ with $m > \beta(n)$ such that, for each $x \in S$, $(p'_{m,\beta(n)})^{-1}(x)$ is not a singleton. Let $\beta(n+1)$ be such an m for the set $S = \{x \in C(X'_{\beta(n)}) \mid |f_n^{-1}(x)| \neq 1\}$. This ensures that the maps f_n and $p'_{\beta(n+1),\beta(n)}$ satisfy the hypothesis of Lemma 8.9 (where $\mathcal{D}_1^{(n)} \cup \{p'_{\beta(n+1),\beta(n)}\}$ plays the role of $\mathcal{D}$ and Z_n plays the role of Z in the application of that lemma). For the ε' we choose $\min\left\{\frac{\varepsilon}{2^{n+2}}, \delta_0\left(\frac{\varepsilon}{2^{n+2}}\right)\right\}$ where δ_0 is a modulus of continuity for $\mathcal{D}_1^{(n)} \cup \{p'_{\beta(n+1),\beta(n)}\}$ (and hence also for $\mathcal{D}_k^{(n)}$ for any $k < n$). The application of Lemma 8.9 produces the map g_n and Lemma 8.10 guarantees that, for any $k < n$, $\mathcal{D}_k^{(n)} \cup \{p'_{\beta(n+1),\beta(n)}, g_n\}$ is $\varepsilon \cdot \left(\frac{1}{2} \cdot \frac{1}{2^{n+1}} + \sum_{m=k}^{n} \frac{1}{2^m}\right)$-commutative. Now to construct f_n, first choose $\alpha(n+1)$ larger than $\alpha(n)$ and large enough so that the maps g_n and $p_{\alpha(n+1),\alpha(n)}$ satisfy the hypothesis of Lemma 8.9.

In preparing to apply Lemma 8.10 we choose $r = \varepsilon/2^{n+2}$ and choose $\varepsilon' = \min\{r, \delta_1(r)\}$, where δ_1 is a modulus of continuity for $\mathcal{D}_1^{(n)} \cup \{p'_{\beta(n+1),\beta(n)}, g_n\}$. Upon applying Lemma 8.9 we obtain the map f_{n+1}. We conclude from Lemma 8.10 that, for $k < n+1$, $\mathcal{D}_k^{(n+1)}$ is $\varepsilon \cdot \left(\frac{1}{2} \cdot \frac{1}{2^{n+1}} + \frac{1}{2} \cdot \frac{1}{2^{n+1}} + \sum_{m=k}^{n} \frac{1}{2^m}\right)$-commutative. □

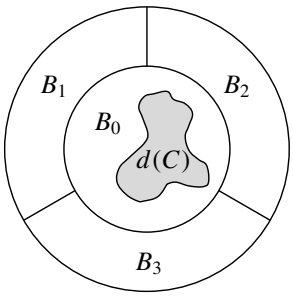
9. Pontryagin cellularity

A compact subset of a Pontryagin surface is *Pontryagin cellular* if it can be expressed as a nested intersection of Pontryagin disks $\mathbb{D}_1, \mathbb{D}_2, \ldots$ where $\mathbb{D}_{i+1} \subset \operatorname{Int} \mathbb{D}_i$ for all i.

Pontryagin cellular subsets of Pontryagin surfaces have some features analogous to those of cellular subsets of genuine surfaces.

Proposition 9.1. *Let C be a compact subset of a rich Pontryagin surface P. Then the decomposition G_C of P whose only nondegenerate element is C is shrinkable if and only if C is Pontryagin cellular in P.*

Proof. The forward implication follows immediately from [Daverman 1986, Proposition 5.12]. For the reverse, given a neighborhood U of C, find a Pontryagin disk $\mathbb{D}$ such that $C \subset \operatorname{Int}\mathbb{D} \subset \mathbb{D} \subset U$. Let $\mathcal{E}$ be a full family of figure-eights for $\mathbb{D}$ and let $d : \mathbb{D} \to B = \mathbb{D}/\mathcal{E}$ denote the quotient map to the resulting disk B. Cover B with a utilitarian web of four disks B_0, B_1, B_2, B_3, as shown below, where $B_0 \subset \mathring{B}$ contains $d(C)$:



Specify a homeomorphism $h : B \to B$ that restricts to the identity on ∂B, that carries B_0 to a disk B_0' whose preimage in P is small, and that sends each of the ∂B_i into $B - d(\mathcal{E})$. Then Corollary 8.5 promises a homeomorphism $H : \mathbb{D} \to \mathbb{D}$ that restricts to the identity on $\partial\mathbb{D}$ and that carries $d^{-1}(B_0)$ to the small set $d^{-1}(B_0')$. Finally, H extends to the rest of P via the identity to give a homeomorphism showing that G_C is shrinkable. □

Proposition 9.2. *A compact subset C of a rich Pontryagin surface P is Pontryagin cellular if and only if C is connected and $P - C$ has an isolated end corresponding to C.*

Proof. The forward implication is routine. For the reverse, note that we can assume P is compact (in view of Proposition 4.3) and connected. Then P/C is both the one-point and end-point compactification of $P - C$. As such, it has a sufficient family $\mathcal{E}_C$ of figure-eights, each of which is contained in $P - C$. Name the quotient map $\psi : P \to P' = P/C$ and the decomposition map $d' : P' \to P'/\mathcal{E}_C$. Given any open subset U of P containing C, one can find a small 2-disk neighborhood B of the point $d'\psi(C)$ in $P'/\mathcal{E}_C$ whose frontier is a simple closed curve missing $d'(\mathcal{E}_C)$, where B satisfies

$$C \subset (d'\psi)^{-1}(\mathring{B}) \subset (d'\psi)^{-1}(B) \subset U.$$

Clearly $(d'\psi)^{-1}(B)$ is a Pontryagin disk. Hence, C is Pontryagin cellular. □

The following observation is used in the proof of the corollary below (other details of which are left to the reader).

Observation. *Suppose X and Y are connected, nice spaces and e is an isolated end of Y. If $f : X \to Y$ is a proper, surjective map which is monotone over some neighborhood of e, then only one end of X is sent to e by f.*

Proof. First note that we can assume without loss that e is the only end of Y (consider $\hat{\hat{f}} | : \hat{\hat{X}} - (\hat{\hat{f}})^{-1}(e) \to \hat{\hat{Y}} - \{e\}$). Supposing that X has more than one end, there exists a neighborhood W of ∞ in X having at least two components which meet ∞. By one-endedness of Y we can find neighborhoods M and N of infinity such that f is monotone over M, $N \subset \mathring{M}$, $\overline{M - N}$ is connected, and $f^{-1}(\overline{M - N}) \subset W$. Hence, $f^{-1}(\overline{M - N})$ is not connected but $f | : f^{-1}(\overline{M - N}) \to \overline{M - N}$ is monotone, thus contradicting the Vietoris–Begle mapping theorem. □

Corollary 9.3. *Let $f : P \to Q$ be a proper, monotone map between Pontryagin surfaces, with P a rich Pontryagin surface. Then each $f^{-1}(q)$, $q \in Q$, is Pontryagin cellular.*

10. Decompositions induced over 1-dimensional subsets and proof of the monotone approximation theorem

The final section of this paper culminates in a proof of the monotone approximation theorem. A key step involves showing how to approximate a given monotone map by one that is injective over certain graphs in the target space.

Proposition 10.1. *Suppose X and Y are compact metrizable spaces and $C_1, C_2, \ldots$ are closed subsets of Y such that, for any surjective monotone map $f : X \to Y$, each of the decompositions $G(C_i)$ induced by f over C_i is shrinkable. Then any such map f can be approximated by a monotone map F that is* 1-1 *over $\bigcup_i C_i$. Moreover, if K is a closed subset of Y such that f is* 1-1 *over K and each of the $G(C_i)$ can be shrunk keeping points over K fixed, then F can be obtained which agrees with f over K.*

Proof. This is a standard Baire category argument. In the complete metric space $\mathscr{S}$ of all surjective, monotone maps $X \to Y$, the collection $O_{j,n}$ of maps f such that $\operatorname{diam} f^{-1}(c) < 1/n$ for all $c \in C_j$ is open, for the usual reasons, and is dense by hypothesis. Any map from the dense subset $\bigcap_{j,n} O_{j,n}$ is 1-1 over $\bigcup_j C_j$.

For the additional control over K, take F as above but form the complete metric subspace of $\mathscr{S}$ consisting of monotone maps $X \to Y$ that agree with f over K. □

Proposition 9.1, Corollary 9.3, and Proposition 10.1 combine to yield:

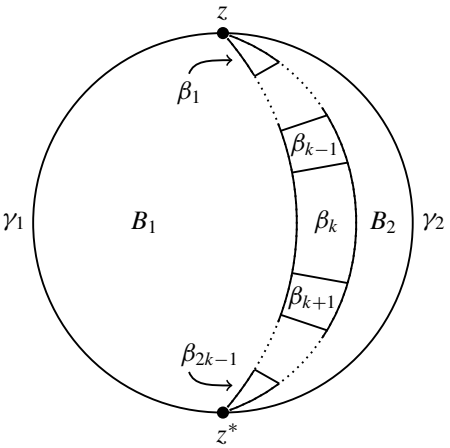
Corollary 10.2. *Let $f : P \to Q$ be a monotone map between rich, closed Pontryagin surfaces, Z a countable subset of Q, and K a closed subset of Q such that f is* 1-1 *over K. Then f can be approximated, arbitrarily closely, by a monotone map F that is* 1-1 *over $Z \cup K$ and agrees with f over K.*

It might be worth mentioning that the next lemma does not show the decomposition under consideration to be shrinkable. The output is merely a homeomorphism that carries decomposition elements to sets of small size — it is not subject to any motion control. A closely related shrinkability result will be established in the subsequent proposition using additional considerations.

Lemma 10.3. *Suppose $\mathbb{D}$ is a Pontryagin disk, $\mathcal{E}$ is a full family of figure-eights for $\mathbb{D}$, $d : \mathbb{D} \to B = \mathbb{D}/\mathcal{E}$ is the decomposition map, and G is a monotone upper semicontinuous decomposition of $\mathbb{D}$ such that the union N_G of all nondegenerate elements of G is a subset of* $\operatorname{Int}\mathbb{D}$ *and the closure of $d(N_G)$ meets ∂B in at most two points. Then for each $\varepsilon > 0$ there exists a homeomorphism $H_\varepsilon : \mathbb{D} \to \mathbb{D}$ such that H_ε restricts to the identity on $\partial\mathbb{D}$ and* $\operatorname{diam} H_\varepsilon(g) < \varepsilon$ *for all $g \in G$.*

Proof. Name the points z, z^* of ∂B containing $\partial B \cap d(ClN_G)$ and let γ_1, γ_2 denote the subarcs of ∂B bounded by these two points.

Identify a disk $B_1 \subset B - \mathring{\gamma}_2$ containing γ_1 such that $\mathring{B}_1$ contains the image under d of every $e \in \mathcal{E}$ with diameter $\varepsilon/6$ or more. Identify another disk $B_2 \subset B$ containing γ_2 that meets B_1 only at the points z and z^*. Cover the rest of B by a chain of 2-cells $\beta_1, \beta_2, \ldots, \beta_{2k-1}$ such that β_i and β_j meet if and only if $|i - j| \leq 1$, the intersection of successive cells β_i and β_{i+1} is an arc in the boundary of each, $z \in \beta_1$ and $z^* \in \beta_{2k-1}$, these β_i together with B_1, B_2 form a utilitarian web for B, and each $d^{-1}(\beta_i)$ has diameter less than $\varepsilon/3$. Furthermore, we can ensure that the 1-skeleton of this utilitarian web avoids the countable set $d(\mathcal{E})$. See:



Now produce an equivalent utilitarian web (equivalent via a homeomorphism fixing ∂B) in B involving 2-cells $B_1', B_2', \beta_1', \beta_2', \ldots, \beta_{2k-1}'$. Here B_1', B_2' should lie very close to γ_1, γ_2, respectively, so as to miss $d(N_G)$. The β_i', except for β_1' and β_{2k-1}', are contained in $\mathring{B}$. In the construction procedure β_k' should be chosen first, and it should meet each of B_1' and B_2' in an arc. Next, β_{k-1}' and β_{k+1}' should be chosen so that any $d(g)$ ($g \in G$) that meets β_k' lives in $\beta_{k-1}' \cup \beta_k' \cup \beta_{k+1}'$. (This is

possible since, by upper semicontinuity and the hypothesis that $d(N_G) \cap \partial B = \varnothing$, elements of G with image near z or z^* are small.) That exposes the general strategy: the 2-cells β'_{k-2} and β'_{k+2} should be chosen, respectively, so that any $d(g)$ that meets β'_{k-1} but not β'_k is contained in $\beta'_{k-2} \cup \beta'_{k-1}$ and so that any $d(g)$ that meets β'_{k+1} but not β'_k is contained in $\beta'_{k+1} \cup \beta'_{k+2}$. The cells $\beta'_{k-3}, \beta'_{k+3}, \ldots$ should be chosen in turn so that, ultimately, any $d(g)$ (g nondegenerate) lies either in $\beta'_{k-1} \cup \beta'_k \cup \beta'_{k+1}$ or in the union $\beta'_{i-1} \cup \beta'_i$ of two successive β'_j. Specify a homeomorphism $h : B \to B$ taking B'_i to B_i and β'_j to β_j and fixing points of ∂B. The homeomorphism H_ε provided by Corollary 8.5 keeps points of $\partial\mathbb{D}$ fixed and shrinks elements of G to size less than ε, since the image of each nondegenerate $g \in G$ lies in some ε diameter set of the form $d^{-1}(\beta_j \cup \beta_{j+1} \cup \beta_{j+2})$. □

Proposition 10.4. *Let $f : P \to Q$ be a degree-one, monotone map between rich, closed Pontryagin surfaces and let A denote any locally separating arc or separating simple closed curve in Q. Then f can be approximated, arbitrarily closely, by monotone maps F that are* 1-1 *over A. Furthermore, the approximations F can be chosen to equal f over any closed subset K of Q such that f is* 1-1 *over K.*

Proof. We will treat only the case in which A is a locally separating arc in Q. The proof for simple closed curves is similar, or can be obtained from the result for arcs plus Proposition 10.1.

By Corollary 10.2 we can approximate f by another monotone map, which we continue to call f, that is 1-1 over a countable, dense subset of A containing ∂A and that agrees with the original f over K. Let $G(A)$ denote the decomposition of P induced by the modified f over A, and let $p : P \to X = P/G(A)$ denote the decomposition map. We show that $G(A)$ is shrinkable fixing points of K.

By the factor theorem (Theorem 3.1) X is a Pontryagin surface and has a full family $\mathcal{E}$ of figure-eights, each of which lives in $X - pf^{-1}(A)$. Let $d : X \to X/\mathcal{E}$ denote the decomposition map associated with the decomposition of X into points and these figure-eights.

Fix $\varepsilon > 0$. Note that d is 1-1 over $A' = dpf^{-1}(A)$. Note also that the closure of each component of $A - K$ has endpoints in $K \cup \partial A$ over which f is one-to-one. It follows easily that only a finite number of components of $A - K$ have preimage under f with diameter at least ε. We let γ denote one of those components. Since we will perform the same operations near each of these components, we assume γ is the only one.

Cover $\gamma' = dpf^{-1}(\gamma) \subset A'$ by a finite collection $B_1, \ldots, B_m$ of 2-cells in the surface $X/\mathcal{E}$. These 2-cells should have pairwise-disjoint interiors and those interiors should miss $dpf^{-1}(K)$, each B_i should meet γ' in an arc whose interior lies in $\operatorname{Int} B_i$, and should be small enough to assure that $d^{-1}(B_i)$ has diameter less than ε. The collection should be arranged so that dp is 1-1 over each $\partial B_i \cap A'$. As

a consequence, each $\mathbb{D}_i = (dp)^{-1}(B_i)$ and each $\mathbb{D}'_i = f(dp)^{-1}(B_i)$ is a Pontryagin disk, with $\partial\mathbb{D}_i$ missing all the nondegenerate elements of $G(A)$.

Now apply Lemma 10.3 m times, using the decomposition induced by $f|\mathbb{D}_i : \mathbb{D}_i \to \mathbb{D}'_i$ on each $\mathbb{D}_i$, to obtain a homeomorphism $H_\varepsilon : P \to P$ that sends each $\mathbb{D}_i$ to itself, restricts to the identity on each $\partial\mathbb{D}_i$ as well as outside $\bigcup_i \mathbb{D}_i$, and sends every nondegenerate $g \in G(A)$ to a set of diameter less than ε. Note that, by construction of the $\mathbb{D}'_i$, f and H_ε are ε-close. Hence, H_ε establishes that $G(A)$ satisfies the shrinkability criterion via shrinking homeomorphisms that reduce to the identity over K.

As in the proof of Corollary 2.5, if $\theta : P \to P/G(A)$ is a homeomorphism very close to p, then $F = fp^{-1}\theta$ is a monotone map close to f which is 1-1 over A and which agrees with f over K. □

Corollary 10.5. *Let $f : P \to Q$ be a degree-one, monotone map between rich, closed Pontryagin surfaces, let $\mathcal{E}$ be a sufficient family of figure-eights for Q, with $d : Q \to S = Q/\mathcal{E}$ the quotient map, and let Γ denote the 1-skeleton of a utilitarian web for S. Then f can be approximated, arbitrarily closely, by a monotone map F that is 1-1 over $d^{-1}(\Gamma)$. Furthermore, if K is a closed subset of Γ such that f is 1-1 over K, then F can be chosen to be equal to f over K.*

Proof. Specify locally separating arcs $A_1, \ldots, A_k$ in Γ covering Γ and then employ Propositions 10.4 and 10.1. □

Proof of the monotone approximation theorem (*Theorem 2.2*). Let $f : P \to Q$ be a degree-one, monotone map between closed, connected, rich Pontryagin surfaces. Given $\varepsilon > 0$, specify a full family $\mathcal{E}_Q$ of figure-eights for Q, and let $\mathcal{E}'_Q$ denote the cofinite subcollection consisting of figure-eights of diameter less than $\varepsilon/4$. Let $d_Q : Q \to S = Q/\mathcal{E}'_Q$ be the associated quotient map to a closed surface S. Find a utilitarian web $W = \{B_1, \ldots, B_m\}$ in S with such small mesh that each $(d_Q)^{-1}(B_i)$ has diameter less than $\varepsilon/2$.

Use Corollary 10.5 to obtain another monotone map $F : P \to Q$ such that F is 1-1 over $d_Q^{-1}(\Gamma)$, where Γ is the 1-skeleton of W, and $\rho(F, f) < \varepsilon/2$.

At this juncture Q has been split into m Pontryagin disks $\mathbb{D}'_i = (d_Q)^{-1}(B_i)$ with pairwise-disjoint interiors, each of diameter less than $\varepsilon/2$. The map F lifts them to Pontryagin disks $\mathbb{D}_i = F^{-1}(\mathbb{D}'_i)$ in P, and F determines monotone maps $F_i = F|\mathbb{D}_i : \mathbb{D}_i \to \mathbb{D}'_i$ that restrict to homeomorphisms $\partial\mathbb{D}_i \to \partial\mathbb{D}'_i$. Corollary 8.2 promises the existence of homeomorphisms $\Phi_i : \mathbb{D}_i \to \mathbb{D}'_i$ that agree with F_i on $\partial\mathbb{D}_i$. By construction of $\mathbb{D}'_i$ each Φ_i is $\varepsilon/2$-close to F_i. Hence, $\Phi = \bigcup_i \Phi_i : \mathbb{D} = \bigcup_i \mathbb{D}_i \to \mathbb{D}' = \bigcup_i \mathbb{D}'_i$ is a homeomorphism which is $\varepsilon/2$-close to F and ε-close to f. □

Theorem 10.6. *Let $f : (\mathbb{D}, \partial\mathbb{D}) \to (\mathbb{D}', \partial\mathbb{D}')$ be a split monotone map between Pontryagin disks and $K \supset \partial\mathbb{D}'$ a closed subset of $\mathbb{D}'$ such that f is 1-1 over K. Then*

f can be approximated, arbitrarily closely, by a homeomorphism $\Phi : \mathbb{D} \to \mathbb{D}'$ such that $\Phi | f^{-1}(K) = f | f^{-1}(K)$.

Proof. The only change to the proof of the monotone approximation theorem required in the Pontryagin disks setting is that in applying Corollary 10.5 one should obtain a monotone map $F : \mathbb{D} \to \mathbb{D}'$ that is 1-1 over the 1-skeleton as before and agrees with f over K. □

References

[Armentrout 1971] S. Armentrout, *Cellular decompositions of 3-manifolds that yield 3-manifolds*, Mem. Amer. Math. Soc. **107**, Amer. Math. Soc., Providence, RI, 1971. MR Zbl

[Bestvina 1988] M. Bestvina, *Characterizing k-dimensional universal Menger compacta*, Mem. Amer. Math. Soc. **380**, 1988. MR Zbl

[Bing 1952] R. H. Bing, "A homeomorphism between the 3-sphere and the sum of two solid horned spheres", *Ann. of Math.* (2) **56** (1952), 354–362. MR Zbl

[Brouwer 1913] L. E. J. Brouwer, "Some remarks on the coherence type η", *KNAW Proc.* **15**:2 (1913), 1256–1263.

[Chapman 1973] T. A. Chapman, "Cell-like mappings of Hilbert cube manifolds: applications to simple homotopy theory", *Bull. Amer. Math. Soc.* **79** (1973), 1286–1291. MR Zbl

[Daverman 1986] R. J. Daverman, *Decompositions of manifolds*, Pure and Appl. Math. **124**, Academic, Orlando, FL, 1986. MR Zbl

[Edwards 1980] R. D. Edwards, "The topology of manifolds and cell-like maps", pp. 111–127 in *Proceedings of the International Congress of Mathematicians* (Helsinki, 1978), vol. I, edited by O. Lehto, Acad. Sci. Fennica, Helsinki, 1980. MR Zbl

[Epstein 1966] D. B. A. Epstein, "The degree of a map", *Proc. London Math. Soc.* (3) **16** (1966), 369–383. MR Zbl

[Fréchet 1910] M. Fréchet, "Les dimensions d'un ensemble abstrait", *Math. Ann.* **68**:2 (1910), 145–168. MR JFM

[Freedman and Quinn 1990] M. H. Freedman and F. Quinn, *Topology of 4-manifolds*, Princeton Math. Series **39**, Princeton Univ., 1990. MR Zbl

[Gu 2017] S. Gu, "Approximating resolutions by cell-like maps with codimension-three point inverses", *Topology Appl.* **232** (2017), 22–28. MR Zbl

[Jakobsche 1991] W. Jakobsche, "Homogeneous cohomology manifolds which are inverse limits", *Fund. Math.* **137**:2 (1991), 81–95. MR Zbl

[Kuratowski and Lacher 1969] K. Kuratowski and R. C. Lacher, "A theorem on the space of monotone mappings", *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.* **17** (1969), 797–800. MR Zbl

[Lacher 1977] R. C. Lacher, "Cell-like mappings and their generalizations", *Bull. Amer. Math. Soc.* **83**:4 (1977), 495–552. MR Zbl

[Mitchell et al. 1992] W. J. R. Mitchell, D. Repovš, and E. V. Ščepin, "On 1-cycles and the finite dimensionality of homology 4-manifolds", *Topology* **31**:3 (1992), 605–623. MR Zbl

[Pontryagin 1930] L. Pontrjagin, "Sur une hypothèse fondamentale de la théorie de la dimension", *C. R. Acad. Sci. Paris* **190** (1930), 1105–1107. Zbl

[Prishlyak and Mischenko 2007] A. O. Prishlyak and K. I. Mischenko, "Classification of noncompact surfaces with boundary", *Methods Funct. Anal. Topology* **13**:1 (2007), 62–66. MR Zbl

[Richards 1963] I. Richards, "On the classification of noncompact surfaces", *Trans. Amer. Math. Soc.* **106** (1963), 259–269. MR Zbl

[Siebenmann 1972] L. C. Siebenmann, "Approximating cellular maps by homeomorphisms", *Topology* **11** (1972), 271–294. MR Zbl

Received June 19, 2017. Revised March 27, 2019.

ROBERT J. DAVERMAN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN
UNITED STATES
rdaverma@utk.edu

THOMAS L. THICKSTUN
DEPARTMENT OF MATHEMATICS
TEXAS STATE UNIVERSITY
SAN MARCOS, TX
UNITED STATES
tt04@txstate.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.133

DENOETHERIANIZING COHEN–MACAULAY RINGS

LÁSZLÓ FUCHS AND BRUCE OLBERDING

We introduce a new class of commutative nonnoetherian rings, called n-subperfect rings, generalizing the almost perfect rings that have been studied recently by Fuchs and Salce. For an integer $n \geq 0$, the ring R is said to be n-subperfect if every maximal regular sequence in R has length n and the total ring of quotients of R/I for any ideal I generated by a regular sequence is a perfect ring in the sense of Bass. We define an extended Cohen–Macaulay ring as a commutative ring R that has noetherian prime spectrum and each localization R_M at a maximal ideal M is $\mathrm{ht}(M)$-subperfect. In the noetherian case, these are precisely the classical Cohen–Macaulay rings. Several relevant properties are proved reminiscent of those shared by Cohen–Macaulay rings.

1. Introduction

The Cohen–Macaulay rings play extremely important roles in most branches of commutative algebra. They have a very rich, fast expanding theory and a wide range of applications where the noetherian hypothesis is essential in most aspects. Cohen–Macaulay rings R are usually defined in one of the following ways:

(a) R is a noetherian ring in which ideals generated by elements of regular sequences are unmixed (i.e., have no embedded primes).

(b) R is a noetherian ring such that the grade (the common length of maximal regular sequences in I) of every proper ideal I equals the height of I.

Several branches of the theory of noetherian rings are known to have natural generalizations to the nonnoetherian case, but there is none that still shares more than a few of the many useful properties of Cohen–Macaulay rings. As a matter of fact, there have been several attempts for generalization, a few reached publication, see [Glaz 1994; Hamilton 2004; Hamilton and Marley 2007; Asgharzadeh and Tousi 2009], but a trade-off for generalization of select properties to quite wide classes of nonnoetherian rings has been the sacrifice of Cohen–Macaulay-like behavior in any

MSC2010: primary 13F99, 13H10; secondary 13C13.
Keywords: Perfect, subperfect, n-subperfect rings, regular sequence, unmixed, Cohen–Macaulay rings.

comprehensive fashion. The noetherian condition has never been replaced by any with direct connection to the noetherian property. We believe that a generalization that is closer to the noetherian condition might allow for new applications and capture more features of Cohen–Macaulay rings than the generalizations in the cited references.

In this note, we are looking for a kind of generalization that is very natural and is as close to Cohen–Macaulay rings as possible, but general enough to be amenable to various applications. We break tradition and choose a different approach: one that does not adhere to any of the classical defining properties. Our strategy is to rephrase the definition to one that does not explicitly require the noetherian condition, to replace the condition that implies the noetherian character by a weaker one, and after doing so, to use the modified definition as the base of generalization.

The following simple characterization of Cohen–Macaulay rings is crucial. To underline its relevance and to draw more attention to this characterization, we include the parallel one for Gorenstein rings though this will not be used in this paper.

Theorem 1.1. *For a commutative noetherian ring R, these are equivalent*:

(i) *R is Cohen–Macaulay*;

(ii) *for every ideal I of R generated by a regular sequence, the quotient ring of R/I is artinian* (i.e., 0-dimensional Cohen–Macaulay).

Similarly, R is Gorenstein if and only if, for every ideal I of R generated by a regular sequence, the quotient ring of R/I is quasi-Frobenius (i.e., 0-dimensional Gorenstein).

Proof. (i) $\Rightarrow$ (ii). Hypothesis (i) implies that the ideal I generated by a regular sequence $x_1, \dots, x_i$ in R is unmixed. Then the quotient ring $Q(R/I)$ of R/I is semilocal noetherian and zero-dimensional, hence artinian.

(ii) $\Rightarrow$ (i). It suffices to prove that if (ii) holds, then every ideal I contains a regular sequence of length $\mathrm{ht}(I)$. We show that if $x_1, \dots, x_t$ is a regular sequence in I and $t < \mathrm{ht}(I)$, then this sequence extends to a regular sequence in I of length $t+1$. Since the quotient ring $Q(R/I)$ is artinian, there are only finitely many minimal prime ideals $P_1, \dots, P_m$ of $(x_1, \dots, x_t)R$, and each element of R not prime to $(x_1, \dots, x_t)R$ is contained in one of the P_j. As $I/(x_1, \dots, x_t)R$ has positive height, $I \not\subseteq P_j$ for any j, so $I \not\subseteq P_1 \cup \cdots \cup P_m$ by prime avoidance. Hence there exists $x_{t+1} \in I$ prime to $(x_1, \dots, x_t)R$, and so $x_1, \dots, x_t, x_{t+1}$ is a regular sequence in I.

To verify the second claim, recall a characterization of Gorenstein rings by Bass [1963, Theorem, p. 9]; it shows that they are Cohen–Macaulay rings such that the primary components of ideals I generated by regular sequences are irreducible, i.e.,

not intersections of two larger ideals. This property is equivalent to saying that the ring R/I (that is now a subdirect product of irreducible rings R/L with the primary components L of I; these L have different prime radicals) has no different, isomorphic simple submodules in its socle. This property of the socle is inherited by the (artinian) quotient ring $Q(R/I)$. By Lam [1999, Theorem 15.27], commutative artinian rings with this property are QF rings. □

Using this observation as the point of departure, we follow our strategy, and want to denoetherianize the artinian property. But nothing is simpler than that: we just replace the descending chain condition on *all* ideals by the descending chain condition on *finitely generated* ideals. We do not stop here, but recall that the descending chain condition on finitely generated ideals is equivalent to the same condition on principal ideals [Björk 1969, Theorem 2], and the latter condition characterizes the *perfect* rings, introduced by Bass [1960]. In conclusion, *we will generalize Cohen–Macaulay rings by replacing "artinian" by "perfect."* More precisely, for an integer $n \geq 0$, we will call a ring R (with maximal regular sequences of lengths n) *n-subperfect* ($n \geq 0$) if the ring of quotients of the ring R/I is perfect for every proper ideal I generated by a regular sequence (and add right away that a 0-subperfect ring is the same as a perfect ring in the sense of Bass). Our nonnoetherian Cohen–Macaulay rings are *the extended Cohen–Macaulay rings:* commutative rings R that have noetherian prime spectra and each localization R_M at a maximal ideal M is $\mathrm{ht}(M)$-subperfect. In our discussion we will concentrate on the n-subperfect case for a fixed $n \geq 0$ (which is more general than the local case).

Asgharzadeh and Tousi [2009] review and compare the various nonnoetherian generalizations of Cohen–Macaulay rings in the literature and add their own variants. In a sense, our generalization lies properly between the classical Cohen–Macaulay rings and their generalizations in the literature, at least as far as zero-dimensional rings are concerned. In fact, a zero-dimensional ring is Cohen–Macaulay if and only if it is artinian, while each of the generalizations listed in [Asgharzadeh and Tousi 2009] includes all zero-dimensional rings in their versions of generalized Cohen–Macaulay rings. In our generalization, in the class of zero-dimensional rings only the perfect rings qualify. (A main difference is in the nilradical: T-nilpotency is properly between being just nil and even nilpotent.) Furthermore, every one-dimensional integral domain is included in all of the previously published generalizations. For the Cohen–Macaulayness however, such domains ought to have artinian factor rings modulo any nonzero ideal, while for our 1-subperfectness these factors are required to be perfect rings. Being closer to the classical version, our generalization is expected to share more analogous properties with Cohen–Macaulay rings than the previous generalizations, yet capture fewer classes of rings. To avoid confusion involving these different generalizations of Cohen–Macaulay

rings, we assume implicitly in what follows that the term "Cohen–Macaulay ring" always designates a *noetherian* Cohen–Macaulay ring.

Let us point out some relevant features of n-subperfect rings that support our claim that this generalization has a number of properties that are fundamental for Cohen–Macaulay rings in the noetherian setting. (Definitions are recalled later. In the following list, n can be any nonnegative integer.)

- A ring R is n-subperfect if and only if its spectrum is noetherian and the localizations R_M are n-subperfect for all maximal ideals M (Corollary 4.6).
- A ring R is n-subperfect if and only if for each regular sequence $x_1, \dots, x_i$ in R $(0 < i \leq n)$, the ring $R/(x_1, \dots, x_i)R$ is $(n-i)$-subperfect (Proposition 3.2).
- An n-subperfect ring is catenary, equidimensional, and of Krull dimension n (Corollary 3.6).
- Direct summand of a direct product of a finite number of n-subperfect rings is n-subperfect (Corollary 4.8).
- A noetherian ring is Cohen–Macaulay if and only if it is an extended Cohen–Macaulay ring as defined above (Corollary 4.4).
- The polynomial ring $R[X_1, \dots, X_n]$, or any of its Veronese subrings, is n-subperfect if and only if R is a perfect ring (Theorems 6.2 and 8.3).
- The grade of a proper ideal I of an n-subperfect ring R (the length t of the longest regular sequence contained in I) is the smallest integer t such that $\operatorname{Ext}^t_R(R/I, R) \neq 0$ (Theorem 3.7).
- If a finite group G operates on an n-subperfect ring R and its order is a unit in R, then the set R^G of ring elements fixed under G is an n-subperfect ring (Corollary 5.2).
- The nilradical N of an n-subperfect ring R is T-nilpotent, and R/N is a Goldie ring (Lemma 2.2, Theorem 5.3).

Our definition leaves ample room for specializations: additional conditions might be added that are not strong enough to enforce the noetherian property, but lead to more pleasant properties of the resulting generalizations (e.g., fixing the injective dimension of the ring as in the Gorenstein case, coherency, or the h-local property might be such a condition). Examples for n-subperfect rings that are not Cohen–Macaulay are abundant; see Section 8.

Our main goal was to get acquainted with the fundamental properties of n-subperfect rings that are analogous to well-known features of Cohen–Macaulay rings. Working in the nonnoetherian situation and in the uncharted territory of subperfect rings meant a challenge in several proofs. We focus our attention to n-subperfectness (i.e., localizations at maximal ideals have the same Krull dimension n — this suffices

to explore the general case) in order to avoid dealing with the complicated general situation corresponding to global Cohen–Macaulay rings that would make the main features less transparent. Occasionally, when it does not obscure the main ideas, we work under the global analogue of Cohen-Macaulay rings; these are the regularly subperfect rings defined in Section 2. (See Corollary 4.4.)

While perhaps less familiar in commutative algebra, perfect rings, the cornerstone of our approach, appear throughout the literature on modules and associative algebras. We review these rings briefly in the next section, but see, for example, [Bass 1960; Lam 2001] for more background. Perfect rings were the leading concept in the theories of almost perfect domains by Bazzoni and Salce [2003] and their generalizations, the almost perfect rings, by Fuchs and Salce [2018]: these rings become one-dimensional Cohen–Macaulay once the noetherian condition is imposed. As an application of our approach, we obtain a well-developed Cohen–Macaulay theory of regular sequences in polynomial rings over perfect rings. Thus, while perfect rings help illuminate the workings of Cohen–Macaulay rings, Cohen–Macaulay rings in turn might help shed new light on the class of perfect rings.

We will also establish a close connection with Goldie rings, another important generalization of noetherian rings. It turns out that n-subperfect rings modulo their T-nilpotent radicals are reduced Goldie rings, so Goldie rings appear naturally in the buildup of our new rings. We have not explored this connection to draw conclusions about the structure of n-subperfect rings. Neither have we investigated as yet the possible denoetherianized Gorenstein version of our generalization where for ideals I of R generated by regular sequences, the quotient rings of R/I are self-injective perfect rings.

2. Definitions and notation

All rings considered here are commutative. We mean by a *perfect ring* a ring over which flat modules are projective. Most of the following characterizations of commutative perfect rings can be found in [Bass 1960, Theorem P; Lam 2001, Theorems 23.20, 23.24]. Recall that a module M is *semiartinian* if every nonzero epic image of M contains a simple submodule.

Lemma 2.1. *The following are equivalent for a commutative ring R*:

(a) *R is a perfect ring.*

(b) *R satisfies the descending chain condition on principal ideals.*

(c) *R is a finite direct product of local rings whose maximal ideals N are T-nilpotent* (i.e., for every sequence $y_1, \ldots, y_n, \ldots$ in N, there is an index m such that $y_1 \cdots y_m = 0$).

(d) *R is semilocal and the localization R_P is perfect for every maximal ideal P.*

(e) *R is semilocal and semiartinian.*

(f) *the finitistic dimension* $\mathrm{Fdim}(R)$ (supremum of finite projective dimensions of R-modules) *is* 0.

(g) *R-modules admit projective covers.* □

We emphasize that "perfect modules and perfect ideals" as they are used, e.g., in [Bruns and Herzog 1998] have nothing to do with perfectness as defined in the preceding lemma.

A ring R is *subperfect* if its total quotient ring $Q(R)$ is perfect, i.e., it is an order in a perfect ring. This is a most essential concept in this paper; it may be viewed as a generalization of the notion of integral domain. All Cohen–Macaulay rings are subperfect. Subperfect rings can be characterized as follows.

Lemma 2.2. *For a commutative ring R, these are equivalent*:

(i) *R is subperfect.*

(ii) *R has only finitely many minimal prime ideals, every zero-divisor in R is contained in a minimal prime ideal, and the nilradical N of R is T-nilpotent.*

(iii) [Gupta 1970] *R satisfies*:
  (a) *the nilradical N of R is T-nilpotent*,
  (b) *R/N is a* (reduced) *Goldie ring* (i.e., it has finite uniform dimension and satisfies the ascending chain condition on annihilators of subsets), *and*
  (c) *R satisfies the regularity condition*: *a regular coset of N can be represented by a regular element of R.* (Moreover, a regular coset of N consists of regular elements of R.)

(iv) [Fuchs and Salce 2018, Lemma 5.5] *The modules over the quotient ring $Q(R)$ are weak-injective as R-modules.*

(v) [Fuchs and Salce 2018, Lemma 5.4] *If M is an R-module of weak dimension ≤ 1, then $Q(R) \otimes_R M$ is a $Q(R)$-projective module.* □

Here an R-module M is said to be *weak-injective* if $\mathrm{Ext}^1_R(A, M) = 0$ for all R-modules A of weak-dimension ≤ 1 [Lee 2006]. The regularity condition with respect to the nilradical was discussed by Small [1966]. His Theorem 2.13 states that a commutative noetherian ring R satisfies this condition if and only if the associated primes of the ideal (0) are the minimal primes of R. (A fourth condition in [Gupta 1970] is automatically satisfied if the ring is commutative.)

It is useful to point out:

Lemma 2.3. *Passing modulo a T-nilpotent ideal preserves subperfectness.*

Proof. If I is a T-nilpotent ideal of a subperfect ring R, then by Lemma 2.2(iii) a regular coset in R/I has a representative that is a regular element of R. Hence it

follows that if Q denotes the quotient ring of R, then Q/I is the quotient ring of R/I, which is a perfect ring. □

An ideal I of the commutative ring R is *subperfect* if $Q(R/I)$ is a perfect ring, i.e., R/I is a subperfect ring. A regular sequence is *subperfect* if the ideal it generates is subperfect. We use the conventions that regular sequences are proper and that the empty sequence is considered a regular sequence. Thus the empty sequence in R is subperfect if and only if R is subperfect.

For several results in Section 3, as well as in later arguments, we work with regular sequences that are not necessarily subperfect. We say a ring R is *regularly subperfect* if each regular sequence of R is subperfect. Thus a ring R is regularly subperfect if and only if for each regular sequence $x_1, \dots, x_i$ in R (including the empty regular sequence), the ring $R/(x_1, \dots, x_i)R$ is subperfect. In particular, a necessary condition for R to be regularly subperfect is that R itself is subperfect. For an integer $n \geq 0$, the ring R is *n-subperfect* if R is regularly subperfect and every maximal regular sequence has length n. As a consequence, R is 0-subperfect if and only if R is perfect. This is because in a 0-subperfect ring every nonunit is a zero-divisor, so $Q(R) = R$.

The 1-subperfect rings are *"almost perfect rings"* (the only difference is that almost perfect rings might have localizations that are perfect rings). These rings have been studied recently; see [Fuchs and Salce 2018; Fuchs 2019]. They were defined as subperfect rings such that each factor ring modulo a regular ideal (i.e., an ideal containing a nonzero-divisor) is a perfect ring.

Lemma 2.4. *Suppose R is a subperfect ring. The following are equivalent*:

(α) *R is almost perfect.*

(β) *Every nonzero torsion R-module contains a simple submodule.*

(γ) *For every regular proper ideal I of R, R/I contains a simple module.*

(δ) *R is h-local and $Q(R)/R$ is semiartinian.* □

Moreover, almost perfect rings have a number of interesting characteristic properties that are new even for Cohen–Macaulay rings of Krull dimension 1. To wit, we mention the following [Fuchs and Salce 2018; Fuchs 2019]. A subperfect ring R is almost perfect if and only if either of the following conditions is satisfied (in (iii) and (iv), envelopes and covers are understood to be part of a genuine cotorsion pair):

(i) All flat R-modules are strongly flat (strongly flat means that it is a summand of a module that is an extension of a free R-module by a direct sum of copies of the ring of quotients Q of R).

(ii) R-modules of weak dimension ≤ 1 are of projective dimension ≤ 1.

(iii) If R is local: every R-module M has a divisible envelope (i.e., a divisible module containing M and being contained in every divisible module that contains M).

(iv) Each R-module M admits a projective dimension 1 cover (i.e., a module of projective dimension ≤ 1 along with a map α to M such that any map from a module of projective dimension ≤ 1 to M factors through α, and no proper summand has this property).

Next we recall some standard terminology. Let R be a ring (commutative), and $R^\times$ the set of regular (nonzero-divisor) elements of R. An element r of $R \setminus I$ is *prime to an ideal* I of R if whenever $s \in R$ with $rs \in I$, then $s \in I$. The set S of elements prime to I is a saturated multiplicatively closed set. The prime ideals of R that contain I and are maximal with respect to not meeting S are the *maximal prime divisors* of I. The prime ideals of R that are minimal with respect to containing I are the *minimal prime divisors* of I. These ideals do not meet S. It follows that the classical ring of quotients $Q(R/I)$ of R/I is R_S/I_S, and the maximal ideals of $Q(R/I)$ are the extensions to $Q(R/I)$ of the maximal prime divisors of I. Similarly, the minimal prime ideals of $Q(R/I)$ are the extensions of the minimal prime divisors of I.

We say an ideal I of the ring R is *unmixed* if every maximal prime divisor of I is also a minimal prime divisor of I; equivalently, $\dim Q(R/I) = 0$. Thus, I is unmixed if and only if every element of R not in a minimal prime divisor of I is prime to I. In the case where R is noetherian, this agrees with the definition of unmixed ideal given by Bruns and Herzog [1998, p. 59]. If R is noetherian, $Q(R/I)$ is semilocal. However, since nonnoetherian rings are our main focus, in our discussions $Q(R/I)$ need not be semilocal without additional assumptions on I.

We say that an ideal I of R is *finitely unmixed* if $Q(R/I)$ is a semilocal zero-dimensional ring. A regular sequence of R is *finitely unmixed* if the ideal it generates is finitely unmixed. Thus every subperfect regular sequence is finitely unmixed, and every finitely unmixed regular sequence is unmixed.

For unexplained terminology we refer to [Matsumura 1986; Bruns and Herzog 1998].

3. Basic properties

Although the focus for most of the article is on n-subperfect rings, in this section we prove several assertions in greater generality.

For an integer $n \geq 0$, say that a ring R is *n-unmixed* if every regular sequence of R extends to a maximal regular sequence of length n that is unmixed. Let $\mathcal{C}$ be a class of zero-dimensional rings. We call a ring R is *n-unmixed* in $\mathcal{C}$ if every regular sequence extends to a maximal regular sequence of length n and for every

regular sequence $x_1, \ldots, x_i$ in R, we have $Q(R/(x_1, \ldots, x_i)R) \in \mathcal{C}$. Thus a ring R is n-subperfect if and only if R is n-unmixed in the class $\mathcal{C}$ of perfect rings.

The property of being n-unmixed in a class $\mathcal{C}$ of zero-dimensional rings can be inductively described, as in the next lemma.

Lemma 3.1. *Let $\mathcal{C}$ be a class of zero-dimensional rings, and $n \geq 1$. A ring R is n-unmixed in $\mathcal{C}$ if and only if for each $0 < i \leq n$ and for each regular sequence $x_1, \ldots, x_i$ in R, the ring $R/(x_1, \ldots, x_i)R$ is $(n-i)$-unmixed in $\mathcal{C}$.*

Proof. Suppose R is n-unmixed in $\mathcal{C}$, and let $0 < i \leq n$. Since R is n-unmixed, every regular sequence that begins with $x_1, \ldots, x_i$ extends to a maximal regular sequence of length n. It follows that every maximal regular sequence in $R/(x_1, \ldots, x_i)R$ has length $n - i$. Also, since every regular sequence in R is unmixed in $\mathcal{C}$, so is every regular sequence in $R/(x_1, \ldots, x_i)R$. Thus, $R/(x_1, \ldots, x_i)R$ is $(n-i)$-unmixed in $\mathcal{C}$.

Conversely, suppose that for each $0 < i \leq n$ and for each regular sequence $x_1, \ldots, x_i$ in R, the ring $R/(x_1, \ldots, x_i)R$ is $(n-i)$-unmixed in $\mathcal{C}$. Let $x_1, \ldots, x_i$ be a regular sequence in R, and let $j \leq i$. Then the zero ideal in $R/(x_1, \ldots, x_j)R$ is by assumption unmixed in $\mathcal{C}$, so $Q(R/(x_1, \ldots, x_j)R) \in \mathcal{C}$. Moreover, since $R/(x_1, \ldots, x_i)R$ is $(n-i)$-unmixed, every maximal regular sequence in this ring has length $n - i$. Thus every extension of $x_1, \ldots, x_i$ $(0 < i \leq n)$ to a maximal regular sequence in R has length n. This proves R is n-unmixed in $\mathcal{C}$. □

Proposition 3.2. *Assume $n \geq 1$. The ring R is n-subperfect if and only if, for each regular sequence $x_1, \ldots, x_i$ $(0 < i \leq n)$ in R, the ring $R/(x_1, \ldots, x_i)R$ is $(n-i)$-subperfect.*

Proof. Apply Lemma 3.1 to the class $\mathcal{C}$ of perfect rings. □

We record the following corollary that also shows how n-perfectness can be defined by induction on n.

Corollary 3.3. *A ring R is n-subperfect $(n \geq 1)$ if and only if it is subperfect and for each regular element $x \in R$, the ring R/xR is $(n-1)$-subperfect.*

Proof. This is an immediate consequence of Proposition 3.2. □

The next lemma follows at once from Lemma 2.3.

Lemma 3.4. *If I is a T-nilpotent ideal of an n-subperfect ring R, then the ring R/I is also n-subperfect.* □

The property of being n-unmixed also has strong consequences for the dimension theory of the ring.

Proposition 3.5. *Suppose $n \geq 0$. If the ring R is n-unmixed, then* $\dim R = n$ *and all maximal chains of prime ideals of R have the same length n.*

Proof. We first prove by induction on n that $\dim R = n$. If $n = 0$, then the empty regular sequence is unmixed, and so $\dim Q(R) = 0$. In this case regular elements are units, therefore we have $R = Q(R)$. Thus, for $n = 0$, $\dim R = 0$ and the claim is clear.

Suppose that $n > 0$ and for each $0 \leq i < n$, every i-unmixed ring has dimension i. We claim that $\dim R = n$. Since R is n-unmixed with $n > 0$, we have $\dim R > 0$. Suppose that $P_0 \subset P_1 \subset \cdots \subset P_m$ is a chain of distinct prime ideals of R with $m > 0$. Since R is n-unmixed with $n > 0$, we have $R \neq Q(R)$ and $\dim Q(R) = 0$. Hence every ideal of R not contained in a minimal prime ideal is regular, so there is a regular $x \in P_1$. By Lemma 3.1, R/xR is $(n-1)$-unmixed. By the induction hypothesis, $\dim R/xR = n - 1$. Since $P_1/xR \subset \cdots \subset P_m/xR$ is a chain of distinct prime ideals of R/xR and $\dim R/xR = n - 1$, we conclude that $m \leq n$. Thus no chain of distinct prime ideals of R has length exceeding n, that is, $\dim R \leq n$. To see that $n \leq \dim R$, use the fact that R has a regular sequence of length n [Kaplansky 1970, Theorem 132]. Therefore, $\dim R = n$.

Next we show that all maximal chains of prime ideals have the same length. The proof is again by induction on n. If $n = 0$, then, as we have established, $\dim R = 0$. In this case the proposition is clear. Let $n > 0$, and suppose the claim holds for all $i < n$. Let $P_0 \subset P_1 \subset \cdots \subset P_k$ and $Q_0 \subset Q_1 \subset \cdots \subset Q_m$ be maximal chains of distinct prime ideals in R. We claim $k = m$. Since the zero ideal of R is unmixed, every nonminimal prime ideal of R is regular. Thus P_1 and Q_1 are regular ideals of R, so there is an $x \in R^\times$ in $P_1 \cap Q_1$. By Lemma 3.1, R/xR is an $(n-1)$-unmixed ring with maximal chains of prime ideals $P_1/xR \subset \cdots \subset P_k/xR$ and $Q_1/xR \subset \cdots \subset Q_m/xR$. By the induction hypothesis on R/xR, we have $k - 1 = m - 1$, thus $k = m$. This means that all chains of maximal length in R have the same length k. It follows that $\dim R = k$, thus $k = \dim R = n$. □

Corollary 3.6. *For every $n \geq 0$, an n-subperfect ring is catenary, equidimensional, and has Krull dimension n.* □

For an ideal I of a ring R, the *I-depth* of R is the smallest positive integer t such that $\operatorname{Ext}^t_R(R/I, R) \neq 0$. If R is noetherian, then the I-depth of R is the length of the longest regular sequence contained in I. Thus a noetherian ring R is Cohen–Macaulay if and only if for each proper ideal I of R, the I-depth of R is equal to the height of I. We show in Theorem 3.7 that this result holds more generally for regularly subperfect rings.

Theorem 3.7. *Let R be a regularly subperfect ring, I a proper ideal of R, and let $n \geq 1$. The following are equivalent*:

(1) *I has height n.*

(2) *Every maximal regular sequence in I has length n.*

(3) *There exists a maximal regular sequence in I of length n.*

(4) $n = \min\{t : \operatorname{Ext}^t_R(R/I, R) \neq 0\}$.

Proof. We first prove the equivalence of (1), (2) and (3). Since the length of a regular sequence in I is at most the height of I, it suffices to show that if $x_1, \ldots, x_t$ is a regular sequence in I such that $t < \operatorname{ht}(I)$, then $x_1, \ldots, x_t$ extends to a regular sequence in I of length $t+1$. Using the fact that $Q(R/(x_1, \ldots, x_t)R)$ is perfect (rather than artinian), we can imitate the proof of Theorem 1.1 to establish the existence of such a regular sequence. Then $x_1, \ldots, x_t, x_{t+1}$ is a regular sequence, and the equivalence of (1), (2) and (3) follows.

To see that (4) implies (2), let $x_1, \ldots, x_t$ be a regular sequence in I, and $J = (x_1, \ldots, x_t)R$. By [Kaplansky 1970, p. 101],

$$\operatorname{Ext}^t_R(R/I, R) \cong \operatorname{Hom}_R(R/I, R/J).$$

Suppose $t < n$. By (4), $\operatorname{Hom}_R(R/I, R/J) = 0$, and hence there does not exist a nonzero element of R/J annihilated by I. Since R is regularly subperfect, R/J is subperfect. By Lemma 2.2 and prime avoidance, I/J is not contained in the set of zero-divisors of R/J, and so $x_1, \ldots, x_t$ extends to a regular sequence of length $t+1$. It follows from (4) that $x_1, \ldots, x_t$ extends to a regular sequence $x_1, \ldots, x_n$ of length n. Since $0 \neq \operatorname{Ext}^n_R(R/I, R) \cong \operatorname{Hom}_R(R/I, R/(x_1, \ldots, x_n)R)$, the image of I in R/J consists of zero-divisors. Thus $x_1, \ldots, x_n$ is a maximal regular sequence in I.

Finally, to see that (3) implies (4), suppose $x_1, \ldots, x_n$ is a maximal regular sequence in I, and let $J = (x_1, \ldots, x_n)R$. (Since I has finite height, such a regular sequence must exist.) Then the image of I in R/J consists of zero-divisors. We first show that there is an element $z \in R \setminus J$ such that $zI \subseteq J$.

By Lemma 2.1(c), $Q := Q(R/J)$ contains orthogonal idempotents $e_1, \ldots, e_n$ such that $1 = e_1 + \cdots + e_n$ and, for each i, $e_i Q$ is a perfect local ring with identity e_i. For a maximal ideal P of Q containing IQ, there is i such that $e_i P$ is the maximal ideal of $e_i Q$. Since the ring $e_i Q$ is semiartinian by Lemma 2.1(e), there exists $y \in Q$ such that $x = e_i y \neq 0$ and $xP = 0$. Hence $xIQ = 0$. From this it follows that we can find $z \in R \setminus J$ such that $zI \subseteq J$.

Define a homomorphism $f : R/I \to R/J$ by $f(r + I) = rz + J$ for all $r \in R$. Then $f \neq 0$, and so by the above isomorphism $\operatorname{Ext}^n_R(R/I, R) \neq 0$. If $t \leq n$ satisfies $\operatorname{Ext}^t_R(R/I/R) \neq 0$, then since (4) implies (3), $x_1, \ldots, x_t$ is a maximal regular sequence in I. By the equivalence of (2) and (3), this yields $t = n$. □

Remark 3.8. From the proof of Theorem 3.7 it is evident that statements (1), (2) and (3) remain equivalent if rather than assuming R is regularly subperfect we assume only that every regular sequence is finitely unmixed.

Corollary 3.9. *Let $n \geq 0$. A ring R is n-subperfect if and only if R is regularly subperfect and each maximal ideal of R has height n.*

Proof. If R is n-subperfect, then each maximal ideal of R has height n by Corollary 3.6. Conversely, if R is regularly subperfect and each maximal ideal has height n, then every maximal regular sequence in R has length n by Theorem 3.7. □

Hamilton and Marley [2007, Definition 4.1] define a ring R to be Cohen–Macaulay if every "strong parameter sequence" on R is a regular sequence. The notion of a strong parameter sequence, which is defined via homology and cohomology of appropriate Koszul complexes, is beyond the scope of our paper. We observe next that regularly subperfect rings are Cohen–Macaulay in this sense.

Corollary 3.10. *Every regularly subperfect ring is Cohen–Macaulay in the sense of Hamilton and Marley.*

Proof. Apply Theorem 3.7 and [Asgharzadeh and Tousi 2009, Theorem 3.4]. □

To verify that a local noetherian ring R of dimension d is Cohen–Macaulay, it is enough to exhibit just one regular sequence of length d. By contrast, the following example shows that in a local domain R of dimension d, the existence of a subperfect regular sequence of length d is not sufficient to guarantee that the domain is d-subperfect.

Example 3.11. Kabele [1971, Example 5] constructs a local domain R having the ring $S = k[\![x, y, z]\!]$ as an integral extension, where k is a field of characteristic 2 with $[k : k^2] = \infty$ and x, y, z are indeterminates for k. The ring R has the property that x, y is not a regular sequence in R, but zR, $(z, x)R$ and $(z, x, y)R$ are distinct prime ideals of R, thus R/zR, $R/(z, x)R$ and $R/(z, x, y)R$ are integral domains, and hence z, x, y is a subperfect regular sequence in R. Moreover, $\dim R = 3$ as S has dimension 3 and is integral over R. Since x, y is not a regular sequence and x is a nonzero-divisor in R, the image of y in R/xR is a zero-divisor. If R is 3-subperfect, then R/xR is subperfect, so y is in a minimal prime ideal P of xR. In this case, Corollary 3.6 implies that $\dim R/P = 2$. Let P' be a prime ideal of S lying over P. S is integral over R, so $\dim S/P' = \dim R/P = 2$ [Kaplansky 1970, Theorem 47, p. 31]. Since S is a catenary domain, this implies $\mathrm{ht}(P') = 1$. However, $(x, y)S$ is a height 2 prime ideal of S contained in P', a contradiction. Therefore, R is not 3-subperfect despite the fact that R has a length 3 maximal regular sequence that is subperfect.

4. Localization and globalization

In this section we consider localization and globalization of the n-subperfect property. In general, issues of localization involving regular sequences are complicated by the fact that a regular sequence in a localization at a prime ideal need not be the

image of a regular sequence in R. However, as we observe in the next lemma, this problem can be circumvented for regularly subperfect rings.

Lemma 4.1. *Let R be a regularly subperfect ring, P a prime ideal of R, and let $x_1, \dots, x_n$ be a regular sequence in R_P. Then there is a regular sequence $y_1, \dots, y_n \in P$ such that*

$$(x_1, \dots, x_i)R_P = (y_1, \dots, y_i)R_P \quad \textit{for each } i = 1, \dots, n.$$

Proof. Let I and J be the ideals of R defined by

$$I = \{r \in R : (\exists s \in R \setminus P)\, rs \in x_1 R\} \quad \text{and} \quad J = \{r \in R : (\exists s \in R \setminus P)\, rs \in x_1 P\}.$$

Then $IR_P = x_1 R_P$ and $JR_P = IPR_P$. Moreover, $J \subset I$ is a proper inclusion, since the image of x_1 in R_P is a nonzero-divisor. $Q(R)$ is zero-dimensional and semilocal, so R has finitely many minimal prime ideals $P_1, \dots, P_m$ such that the set of zero-divisors in R is $P_1 \cup \cdots \cup P_m$. Since the image of x_1 in R_P is a nonzero-divisor, $I \not\subseteq P_j$ for any j. By prime avoidance, there is $y_1 \in I$ such that $y_1 \notin J \cup P_1 \cup \cdots \cup P_m$. Since IR_P is a principal ideal and the image of y_1 in R_P is not in JR_P, Nakayama's lemma implies $x_1 R_P = IR_P = y_1 R_P$. By the choice of y_1, we have $y_1 \in R^\times$.

Now suppose $1 < t \leq n$ and there is a regular sequence $y_1, \dots, y_{t-1}$ with $(x_1, \dots, x_i)R_P = (y_1, \dots, y_i)R_P$ for each $1 \leq i \leq t-1$. Then $Q(R/(y_1, \dots, y_{t-1})R)$ is semilocal and zero-dimensional, so repeating the argument from the first paragraph for the ring $R/(y_1, \dots, y_{t-1})R$ yields $y_t \in P$ such that $y_1, \dots, y_{t-1}, y_t$ is a regular sequence in P and $(y_1, \dots, y_{t-1}, y_t)R_P = (x_1, \dots, x_{t-1}, x_t)R_P$. □

Theorem 4.2. *Let R be a regularly subperfect ring. For each prime ideal P of R, the ring R_P is regularly subperfect.*

Proof. Let P be a prime ideal of R. Since $Q(R)$ is zero-dimensional, $Q(R_P) = Q(R)_{R \setminus P}$; see [Lipman 1965, Proposition 1 and Corollary 1]. Thus $Q(R_P)$ is perfect since $Q(R)$ is, and so R_P is subperfect. It follows that the localization of a regularly subperfect ring at a prime ideal has the property that the empty regular sequence is subperfect.

We now prove the theorem by induction on the length of regular sequences in R_P. Let $n > 0$, and suppose that for every regularly subperfect ring S and prime ideal L of S, every regular sequence of length $< n$ in S_L is subperfect. Let $x_1, \dots, x_n$ be a sequence in R whose image in R_P is a regular sequence. By Lemma 4.1 there is $y \in R^\times$ such that $x_1 R_P = yR_P$. Since R/yR is regularly subperfect and the image of the sequence $x_2, \dots, x_n$ in $R_P/x_1 R_P = R_P/yR_P$ is a regular sequence of length $n-1$, the induction hypothesis implies that $R_P/x_1 R_P$ is regularly subperfect. Therefore, the image of the sequence $x_2, \dots, x_n$ in $R_P/x_1 R_P$ is a subperfect regular sequence, and hence so is the image of the sequence $x_1, x_2, \dots, x_n$ in R_P. □

Corollary 4.3. *Let R be a regularly subperfect ring. If P is a prime ideal of finite height n, then R_P is n-subperfect.*

Proof. This follows from Theorems 3.7 and 4.2. □

Corollary 4.4. *The following are equivalent for a noetherian ring R.*

(1) *R is Cohen–Macaulay.*

(2) *R is regularly subperfect.*

(3) *R_M is $\mathrm{ht}(M)$-subperfect for each maximal ideal M of R.*

Proof. To see that (1) implies (2), let $x_1, \ldots, x_n$ be a regular sequence in R. By the unmixedness theorem [Bruns and Herzog 1998, Theorem 2.1.6, p. 59], $x_1, \ldots, x_n$ is unmixed (as is the empty regular sequence). Since R is noetherian, the zero-dimensional ring $Q(R/(x_1, \ldots, x_n R)$ is semilocal, hence artinian, hence perfect. Consequently, the sequence $x_1, \ldots, x_n$ is subperfect.

That (2) implies (3) follows from Corollary 4.3. That (3) implies (1) is clear. □

A topological space is *noetherian* if its open sets satisfy the ascending chain condition. It follows that every closed subset of a noetherian space is a union of finitely many irreducible components. Thus, if R is a ring for which $\mathrm{Spec}(R)$ is noetherian, then each proper ideal of R has but finitely many minimal prime divisors.

Theorem 4.5. *Let R be a ring of finite Krull dimension. Then R is regularly subperfect if and only if $\mathrm{Spec}(R)$ is noetherian and R_M is regularly subperfect for each maximal ideal M of R.*

Proof. Suppose R is regularly subperfect. By Theorem 4.2, R_M is regularly subperfect for each maximal ideal M of R. The proof that $\mathrm{Spec}(R)$ is noetherian is by induction on $\dim R$. If $\dim R = 0$, then R is subperfect, hence perfect, since the ideal (0) of R is generated by the empty regular sequence; thus $\mathrm{Spec}(R)$ is noetherian in this case. Suppose $\dim R > 0$, and for each $0 \leq k < \dim R$ every k-dimensional regularly subperfect ring has a noetherian spectrum. Since R is subperfect, R has only finitely many minimal prime ideals $P_1, \ldots, P_m$. Thus $\mathrm{Spec}(R)$ is a finite union of the closed sets consisting of the prime ideals containing a given minimal prime ideal P_j. To prove that $\mathrm{Spec}(R)$ is noetherian, we need only verify that each of the spaces $\mathrm{Spec}(R/P_j)$ is noetherian. A space is noetherian if and only if it satisfies the descending chain condition on closed sets, therefore we need only prove that every proper closed subset of $\mathrm{Spec}(R/P_j)$ is noetherian. Every proper closed subset of $\mathrm{Spec}(R/P_j)$ is homeomorphic to a subspace of $\mathrm{Spec}(R/(rR + P_j))$ for some $r \in R \setminus P_j$. Therefore, we treat only spectra of rings of the latter form.

Suppose $r \in R \setminus P_j$ for some $1 \leq j \leq m$, and choose $p_j \in R$ such that p_j is contained in exactly the minimal prime ideals of R that do not contain r. (This is possible by prime avoidance and the fact that there are only finitely many minimal prime ideals of R.) In particular, $p_j \in P_j$. Evidently, $r + p_j \notin P_1 \cup \cdots \cup P_m$, so that $r + p_j \in R^\times$. Thus $R/(r + p_j)R$ inherits from R the property that each regular sequence is subperfect. By the induction hypothesis, $\operatorname{Spec}(R/(r + p_j)R)$ is a noetherian space. As a subspace of a noetherian space, $\operatorname{Spec}(R/(rR + P_j))$ is noetherian. This completes the proof that $\operatorname{Spec}(R)$ is a noetherian space.

Conversely, suppose $\operatorname{Spec}(R)$ is noetherian, and R_M is regularly subperfect for each maximal ideal M of R. Let $x_1, \ldots, x_t$ be a (possibly empty) regular sequence in R, and let $I = (x_1, \ldots, x_t)R$. For each maximal ideal M containing I, the images of $x_1, \ldots, x_t$ in R_M form a regular sequence, so R_M/IR_M is subperfect by assumption. We claim that $Q(R/I)$ is zero-dimensional. Let $r, s \in R$ such that $rs \in I$ and r is not contained in any minimal prime ideal of I. It suffices to show that $s \in I$. If M is any maximal ideal of R containing I, then since R_M/IR_M is subperfect and rR_M is not a subset of any minimal prime ideal of IR_M, we have $sR_M \subseteq IR_M$. Since this is true for each maximal ideal M containing I, we conclude that $s \in I$. This proves that every zero-divisor in R/I is contained in a minimal prime ideal of R/I. Therefore, $Q(R/I)$ is zero-dimensional.

Since $\operatorname{Spec}(R)$ is noetherian, I has only finitely many minimal prime ideals $P_1, \ldots, P_m$, so $Q(R/I)$ is also semilocal. For each j, R_{P_j}/IR_{P_j} is T-nilpotent, so it follows that $Q(R/I)$ has T-nilpotent nilradical, and hence $Q(R/I)$ is perfect. This proves that every regular sequence in R (including the empty sequence) is subperfect. □

Corollary 4.6. *Assume $n \geq 0$. A ring R is n-subperfect if and only if* $\operatorname{Spec}(R)$ *is noetherian and R_M is n-subperfect for each maximal ideal M of R.*

Proof. Apply Corollary 3.9 and Theorem 4.5. □

Remark 4.7. The proofs of Lemma 4.1 and Theorems 4.2 and 4.5 show that in the hypotheses of these results the property of being regularly subperfect can be replaced by the more general condition that every regular sequence is finitely unmixed.

We record an immediate consequence of Corollary 4.6 along with the obvious statement on the behavior of n-subperfectness under passing to summands.

Corollary 4.8. *Every summand of a direct product of a finite number of n-subperfect rings is n-subperfect.* □

5. More on n-subperfect rings

We would like to point out several important properties that are shared by n-subperfect rings with Cohen–Macaulay rings. The first of these properties, proved

by Hochster and Eagan [1971] for Cohen–Macaulay rings, concerns descent of the n-subperfect property to module direct summands and to rings of invariants of n-subperfect rings.

Theorem 5.1. *Let R be an n-subperfect ring for some $n \geq 0$. If S is a subring of R such that R is integral over S and S is a direct summand of R as an S-module, then S is n-subperfect.*

Proof. First we claim that S is subperfect. Every minimal prime ideal of S is contracted from a minimal prime ideal of R. Since R is subperfect, there are but finitely many minimal prime ideals of R, so there are only finitely many minimal prime ideals of S. Moreover, every zero-divisor in R is an element of a minimal prime ideal of the subperfect ring R, so the same holds for S. Since the nilradical of S is contained in that of R, it is T-nilpotent. Consequently, S is subperfect.

The proof proceeds now by induction on n. Suppose $n = 0$, so that R is perfect. Then $\dim R = 0$, and since R is integral over S, we have $\dim S = 0$. Since S is subperfect, this implies S is perfect, i.e., 0-subperfect.

Now suppose $n > 0$ and that the claim holds for $n - 1$. If every nonzero-divisor of S were a unit, then since S is subperfect, we would have $\dim S = 0$. R is integral over S, whence $\dim R = 0$ would follow. However, R is n-subperfect, so $\dim R = n > 0$ by Corollary 3.6. Therefore, there exist regular sequences in S of length > 0. Let $s \in S^\times$ be a nonunit in S. Since S is a summand of R, it follows that $sR \cap S = sS$; see [Bruns and Herzog 1998, Lemma 6.4.4]. Thus S/sS can be viewed as a direct summand of R/sR. Moreover, R/sR is integral over S/sS.

To see that $s \in R^\times$, suppose to the contrary that s is a zero-divisor in R. Since R is subperfect, s is contained in a minimal prime ideal P_0 of R. By Corollary 3.9, there is a chain of distinct prime ideals $P_0 \subset P_1 \subset \cdots \subset P_n$, with P_n a maximal ideal of R. Since R is integral over S, the chain $P_0 \cap S \subset P_1 \cap S \subset \cdots \subset P_n \cap S$ has length n. Again since R is integral over S, each chain of prime ideals of S has a chain of prime ideals in R lying over it. Therefore, Corollary 3.9 implies that the length of the longest chain of prime ideals in S is n. Consequently, $P_0 \cap S$ is a minimal prime ideal of R. However, $s \in P_0 \cap S$ and $s \in S^\times$, a contradiction that implies $s \in R^\times$.

In view of $s \in R^\times$, we have R/sR is $(n-1)$-subperfect by Proposition 3.2. By the induction hypothesis, S/sS is $(n-1)$-subperfect. Since this is the case for all nonunits $s \in S^\times$, Corollary 3.3 implies S is n-subperfect, completing the induction. □

Corollary 5.2. *Assume G is a finite group acting on an n-subperfect ring R, and the order of G is a unit in R. Then the set of invariants,*

$$R^G = \{r \in R : \ g(r) = r \text{ for all } g \in G\},$$

is again an n-subperfect ring.

Proof. As in [Bruns and Herzog 1998, pp. 281–283], the hypotheses imply that R^G is a module direct summand of R and R is integral over R^G. Thus we may apply Theorem 5.1 to obtain the corollary. □

Lemma 2.2 makes it possible to get more information on n-subperfect rings once we know more about Goldie rings.

A commutative reduced Goldie ring R is an order in a semisimple ring Q that is the direct product of fields Q_j,

$$Q = Q_1 \times \cdots \times Q_m$$

(see [Lam 1999, Proposition 11.22]). If $X_j = \sum_{i \neq j} Q_i$, then $P_j = X_j \cap R$ $(j = 1, \ldots, m)$ is the set of minimal primes of R. Furthermore, each R/P_j is an integral domain with Q_j as quotient field. Recall that orders R, R' in a ring Q are *equivalent* if $qR \subseteq R'$ and $q'R' \subseteq R$ for some units $q, q' \in Q$.

Theorem 5.3. *A reduced n-subperfect ring R is a Goldie ring. It is a subdirect product of a finite number of integral domains of Krull dimension n. This subdirect product is equivalent to the direct product of the components.*

Proof. Assume R is reduced and n-subperfect; in view of Lemma 2.2, it is a Goldie ring. It has but a finite number of minimal prime ideals $P_1, \ldots, P_m$. From $\bigcap_j P_j = 0$ it follows that R is a subdirect product of the integral domains $D_j = R/P_j$ (with quotient fields Q_j). It is clear that $\dim D_j = n$ for each j.

Choose elements x_j $(j = 0, \ldots, m)$ such that $x_j \in P_i$ for all $i \neq j$, but $x_j \notin P_j$. Then $x = \sum_j x_j \in R$ is a regular element, as it is not contained in any P_j. Therefore,

$$x = (x_1 + P_1, \ldots, x_m + P_m) \in D_1 \oplus \cdots \oplus D_m$$

is a unit in Q. Hence we conclude that R and $R' = D_1 \oplus \cdots \oplus D_m$ are equivalent orders in Q. □

We observe that Theorem 5.3 holds also for the factor ring R/N of an n-subperfect ring R modulo its nilradical N, though R/N need not be n-subperfect. Note that this factor ring is restricted in size inasmuch as R/N must have finite uniform dimension. On the other hand, Example 8.2 will show that the nilradicals of n-subperfect rings can have arbitrarily large cardinalities.

We have failed to establish a stronger result in the preceding theorem (viz. that the domains D_j are also n-subperfect), because passing modulo a minimal prime ideal, regular sequences do not map in general upon regular sequences, though the converse is true for all regularly subperfect rings as is shown by:

Lemma 5.4. *Let R be a regularly subperfect ring, and let P be a minimal prime ideal of R. Then for every regular sequence $y_1, \ldots, y_t$ in $S = R/P$, there is a regular sequence $x_1, \ldots, x_t \in R$ such that $(x_1, \ldots, x_i)S = (y_1, \ldots, y_i)S$ for all $i \leq t$.*

Proof. The proof is by induction on the length of the regular sequence. The claim is clearly true for the empty regular sequence. Suppose that $t \geq 0$ and the claim is true for all regular sequences in S of length t. Let $y_1, \ldots, y_t, y_{t+1}$ be a regular sequence in S. Then there is a regular sequence $x_1, \ldots, x_t$ in R such that $(x_1, \ldots, x_t)S = (y_1, \ldots, y_t)S$. Since $R/(x_1, \ldots, x_t)R$ is subperfect, $(x_1, \ldots, x_t)R$ has but a finite number of minimal prime ideals $L_1, \ldots, L_k$. Let $z_{t+1} \in R$ such that $z_{t+1} + P = y_{t+1}$. We observe that $P + z_{t+1}R \not\subseteq L_i$ for any i. Indeed, if $P \subseteq L_i$ for some i, then L_i is a minimal prime ideal of $(x_1, \ldots, x_t)R + P$. In this case, since $y_1, \ldots, y_{t+1}$ is a regular sequence in S and $(x_1, \ldots, x_t)S = (y_1, \ldots, y_t)S$, it is impossible to have $y_{t+1} \in L_i/P$. Thus $z_{t+1} \notin L_i$ which shows that $P + z_{t+1}R \not\subseteq L_i$ for every i. By a version of prime avoidance [Kaplansky 1970, Theorem 124], this implies there is $p \in P$ such that $z_{t+1} - p \notin L_i$ for each i. Since $L_1, \ldots, L_k$ are the minimal prime ideals of $(x_1, \ldots, x_t)R$ and $R/(x_1, \ldots, x_t)R$ is subperfect, it follows that $x_1, \ldots, x_t, x_{t+1}$ with $x_{t+1} = z_{t+1} - p$ is a regular sequence in R such that $(x_1, \ldots, x_t, x_{t+1})S = (y_1, \ldots, y_{t+1})S$. This completes the induction and shows that every ideal of S generated by a regular sequence is the image of an ideal of R that is generated by a regular sequence. □

The next theorem shows that for regularly subperfect rings, ideals of the principal class (i.e., ideals I generated by $\mathrm{ht}(I)$ elements) behave like ideals in Cohen–Macaulay rings.

Theorem 5.5. *Let R be a regularly subperfect ring, and let I be an ideal of R generated by t elements. The following are equivalent*:

(1) *I has height t.*

(2) *I has height at least t.*

(3) *I is generated by a regular sequence of length t.*

Proof. That (1) implies (2) is clear, and that (3) implies (1) follows from Theorem 3.7. To see that (2) implies (3), suppose $\mathrm{ht}(I) \geq t$. If $\mathrm{ht}(I) = 0$, then I is generated by the empty regular sequence. The proof now proceeds by induction on $\mathrm{ht}(I)$. Suppose that in a regularly subperfect ring, every ideal $I = (x_1, \ldots, x_t)R$ of height at least $\mathrm{ht}(I)-1$ generated by $\mathrm{ht}(I)-1$ elements is generated by a regular sequence of length $\mathrm{ht}(I) - 1$. As a subperfect ring, R admits only finitely many minimal prime ideals $P_1, \ldots, P_m$. Prime avoidance and the fact that $\mathrm{ht}(I) > 0$ imply that $I \not\subseteq P_1 \cup \cdots \cup P_m$. By [Kaplansky 1970, Theorem 124, p. 90], there exist $r_2, \ldots, r_t \in R$ such that $x := x_1 + r_2x_2 + \cdots + r_tx_t \notin P_1 \cup \cdots \cup P_m$. Since R is subperfect, $x \in R^\times$. Moreover, $I = (x, x_2, \ldots, x_t)R$. In order to apply the induction hypothesis, we consider next the ring R/xR.

Let P be a minimal prime ideal of I such that $\mathrm{ht}(P) = \mathrm{ht}(I)$. By Theorem 4.2, R_P is $\mathrm{ht}(I)$-subperfect, so Proposition 3.2 implies R_P/xR_P is $(\mathrm{ht}(I)-1)$-subperfect.

By Corollary 3.6, $\dim R_P/xR_P = \mathrm{ht}(I) - 1$, and so P/xR has height $\mathrm{ht}(I) - 1$ in R/xR. Consequently, P/xR is a minimal prime ideal of I/xR of height $\mathrm{ht}(I)-1$ in R/xR. Thus I/xR is an ideal of R/xR that is generated by $t-1$ elements and has height at least $\mathrm{ht}(I) - 1$. By the induction hypothesis, I/xR is generated by a regular sequence in R of length $t-1$. Thus I is generated by a regular sequence of length t. This proves that every ideal of R of height at least t generated by t elements is generated by a regular sequence of length t. Consequently, (2) implies (3). □

6. Polynomial rings

We consider next polynomial rings $S = R[X_1, \ldots, X_n]$ over a perfect ring R. The proof of Theorem 6.2, which shows such rings are n-subperfect, depends on the following lemma.

Lemma 6.1. *Let S be a finitely generated algebra over a perfect ring R. For each proper ideal I of S, the nilradical of S/I is T-nilpotent. If also* $\dim Q(S/I) = 0$, *then S/I is subperfect.*

Proof. Let I be a proper ideal of S. Then the nilradical of S/I is $\sqrt{I}/I$, so to show that this nilradical is T-nilpotent, it suffices to show that for all $a_1, a_2, a_3, \ldots \in \sqrt{I}$, there exists $m > 0$ such that $a_1 a_2 \cdots a_m \in I$. We claim first that there is $k > 0$ such that $(\sqrt{I})^k \subseteq I + JS$, where J denotes the Jacobson radical of R. Since R/J is an artinian ring (it is a product of finitely many fields) and S/JS is a finitely generated R/J-algebra, the ring S/JS is noetherian. Thus the image of the ideal $\sqrt{I}$ in S/JS is finitely generated. Letting $f_1, \ldots, f_t \in \sqrt{I}$ such that $\sqrt{I} = (f_1, \ldots, f_t)S + JS$, and choosing $k > 0$ such that $(f_1, \ldots, f_t)^k S \subseteq I$, we obtain $(\sqrt{I})^k \subseteq I + JS$.

For each $i \geq 0$, we have $a_{ik+1}a_{ik+2} \cdots a_{ik+k} \in I + JS$, and so there is a finitely generated ideal $A_i \subseteq J$ such that $a_{ik+1}a_{ik+2} \cdots a_{ik+k} \in I + A_i S$. As a perfect ring, R satisfies the descending chain condition on finitely generated ideals [Björk 1969, Theorem 2], thus there is $t > 0$ such that $A_1 A_2 \cdots A_t = A_1 A_2 \cdots A_{t+1}$. Since $A_{t+1} \subseteq J$, Nakayama's lemma implies $A_1 A_2 \cdots A_t = 0$. It follows that

$$a_1 a_2 \cdots a_{tk+k} \in (I + A_0 S)(I + A_1 S) \cdots (I + A_t S) \subseteq I,$$

which proves the first assertion.

Now suppose $\dim Q(S/I) = 0$. Since R is perfect, $\mathrm{Spec}(R)$ is a finite, hence noetherian, space. As a finitely generated algebra over a ring with noetherian prime spectrum, S also has noetherian prime spectrum [Ohm and Pendleton 1968, Theorem 2.5]. Hence I has but finitely many minimal prime divisors, and so, since $Q(S/I)$ is zero-dimensional, it follows that $Q(S/I)$ is semilocal. The nilradical of $Q(S/I)$ is T-nilpotent as it is extended from the T-nilpotent nilradical of S/I; hence $Q(S/I)$ is perfect. □

We now prove the main theorem of this section. Statement (4) of Theorem 6.2, which is a byproduct of our arguments involving polynomial rings, can be viewed as a characterization of a perfect ring in terms of its multiplicative lattice of ideals.

Theorem 6.2. *Let R denote a commutative ring, and let $X_1, \dots, X_n$ $(n \geq 1)$ be indeterminates for R. Then the following are equivalent*:

(1) *R is perfect.*

(2) *$R[X_1, \dots, X_n]$ is n-subperfect.*

(3) *R is semilocal zero-dimensional and $R[X_1, \dots, X_n]$ is subperfect.*

(4) *R is semilocal zero-dimensional and for each sequence $\{I_i\}_{i=1}^{\infty}$ of finitely generated subideals of the Jacobson radical J of R there exists $k > 0$ such that $I_1 I_2 \cdots I_k = 0$.*

Proof. Let $S = R[X_1, \dots, X_n]$, and let J denote the Jacobson radical (= the nilradical) of R.

(1) $\Rightarrow$ (4). Perfect rings are semilocal and zero-dimensional. Let $\{I_i\}_{i=1}^{\infty}$ be a sequence of finitely generated subideals of J. Since R is perfect, R satisfies the descending chain condition on finitely generated ideals [Björk 1969, Theorem 2], thus there is $k > 0$ such that $I_1 I_2 \cdots I_k = I_1 I_2 \cdots I_{k+1}$. Since $I_{k+1} \subseteq J$, Nakayama's lemma implies $I_1 I_2 \cdots I_k = 0$.

(4) $\Rightarrow$ (3). Let $f_1/g_1, f_2/g_2, \dots$ be elements of the nilradical of $Q(S)$, where each $f_i \in S$ and each g_i is a nonzero-divisor in S. Then $f_1, f_2, \dots$ are in the nilradical of S, which, since S is a polynomial ring, is the extension JS of the nilradical J of R to S. The ideal I_i generated by the coefficients occurring in f_i is contained in the nilradical of R, so by assumption, there is $k > 0$ such that $I_1 I_2 \cdots I_k = 0$. Since $f_1 f_2 \cdots f_k \in I_1 I_2 \cdots I_k S$, we have $f_1 f_2 \cdots f_k = 0$, thus the nilradical of $Q(S)$ is T-nilpotent. Furthermore, since R is zero-dimensional, so is $Q(S)$ by [Arapović 1983, Proposition 8]. Each prime ideal L in $Q(S)$ contracts to one of the prime ideals P in R. Since $PQ(S) \subseteq L$ is a prime ideal of $Q(S)$ and $Q(S)$ is zero-dimensional, it follows that $PQ(S) = L$. Therefore, since R is semilocal, so is $Q(S)$. This shows that $Q(S)$ is a zero-dimensional semilocal ring with T-nilpotent nilradical; i.e., $Q(S)$ is perfect.

(3) $\Rightarrow$ (2). Suppose S is subperfect. Let $f_1, \dots, f_t$ be a regular sequence in S, and let $I = (f_1, \dots, f_t)S$. Now R is zero-dimensional and semilocal and I is generated by a regular sequence, therefore — as it is shown in [Olberding 2019] — the ring $Q(S/I)$ is also zero-dimensional and semilocal. By Lemma 6.1, $Q(S/I)$ is a perfect ring, establishing the n-subperfectness of S.

(2) $\Rightarrow$ (1). Since S is n-subperfect and $X_1, \ldots, X_n$ is a maximal regular sequence of S, $S/(X_1, \ldots, X_n)R$ is a perfect ring. As a homomorphic image of this ring, R is perfect. □

Let us point out that Coleman and Enochs [1971] prove that if the polynomial rings $R[X]$ and $R'[Y]$ with single indeterminates are isomorphic, and if R is a perfect ring, then $R \cong R'$. It is an open problem if this holds for more indeterminates.

7. The finitistic dimension

The close relation of n-subperfect rings to Goldie rings makes it possible to derive several interesting properties of n-subperfect rings. For details we refer to the literature on Goldie rings, e.g., [Goodearl and Warfield 1989]. As an example we mention that the ring of quotients of a reduced n-subperfect ring is its injective hull.

In view of [Sandomierski 1973], we are able to obtain interesting results on the homological dimensions of n-subperfect rings. We show that in calculating the projective (p.d.), injective (i.d.) and weak (w.d.) dimensions of modules over an n-subperfect ring, only the "Goldie part" of the ring counts (see Lemma 2.2).

Let R be an n-subperfect ring with minimal prime ideals $P_1, \ldots, P_m$. Then $N = P_1 \cap \cdots \cap P_m$ is the nilradical of R; it is T-nilpotent. By Theorem 5.3, R/N is a subdirect product of n-dimensional integral domains $D_j = R/P_j$ $(j = 1, \ldots, m)$. In the next theorem, D_j-modules are also regarded as R-modules in the natural way.

Theorem 7.1. *Let R denote an n-subperfect ring, and let D_j be as before. Then an R-module M satisfies* $\text{p.d.}_R M \le k$ $(k \ge 0)$ *if and only if* $\text{Ext}_R^{k+1}(M, X) = 0$ *for all D_j-modules X for each $j = 1, \ldots, m$.*

Proof. See Theorem 5.3 in [Sandomierski 1973]. □

Theorem 7.2. *Let R be an n-subperfect ring, and $R^* = R/N$. Then for an R-module M we have for any $k \ge 0$:*

(a) $\text{p.d.}_R M \le k$ *if and only if* $\text{Ext}_R^{k+1}(M, X) = 0$ *for all R^*-modules X.*

(b) $\text{i.d.}_R M \le k$ *if and only if* $\text{Ext}_R^{k+1}(R/L, M) = 0$ *for all ideals L containing N.*

(c) $\text{w.d.}_R M \le k$ *if and only if* $\text{Tor}_{k+1}^R(R/L, M) = 0$ *for all ideals L containing N.*

Proof. See Theorems 5.2, 3.2, and 4.2, respectively, in [Sandomierski 1973]. □

Also, [Sandomierski 1973, Proposition 5.4] shows that for a flat R-module F, $\text{p.d.}_R F$ can be calculated as the maximum of the D_j-projective dimensions of the flat D_j-modules $F \otimes_R D_j$, taken for all j.

We would like to have information about the finitistic dimension $\text{Fdim}(R)$ of an n-subperfect ring R. An estimate is given by [Sandomierski 1973, Corollary 1, Section 2] which we cite using the same notation as above.

Theorem 7.3. *For an n-subperfect ring R and for the integral domains D_j we have the inequality*

$$\operatorname{Fdim}(R) \le \max_j\{\text{p.d.}_R D_j + \operatorname{Fdim}(D_j)\}. \qquad \square$$

We recall (see, e.g., [Jensen 1972, Remarque, p. 44]) that for a Cohen–Macaulay ring R, the finitistic dimension $\operatorname{Fdim}(R)$ is equal either to d or to $d+1$ where $d = \dim R$. For n-subperfect rings we do not have such a precise estimate, but we still have some information, see Theorem 7.5.

In the balance of this section, we will use the notation $\mathcal{P}_n(R)$ for the class of R-modules whose projective dimensions are $\le n$, and $\mathcal{F}_n(R)$ for the class of modules of weak dimensions $\le n$. We concentrate on the class $\mathcal{F}_1(R)$ which is more relevant to subperfectness than the class $\mathcal{F}_0(R)$ of flat modules; see, e.g., Lemma 2.2(v).

Next, we verify a lemma (note that $\bar{R}$-modules may be viewed as R-modules).

Lemma 7.4. *Let R be any ring and $\bar{R} = R/rR$ with $r \in R^\times$ a nonunit. Then*

(1) *if $\bar{R}$ is subperfect, then $\mathcal{F}_1(\bar{R}) \subseteq \mathcal{F}_1(R)$;*

(2) *if both R and $\bar{R}$ are subperfect, then $\mathcal{F}_1(R) \subseteq \mathcal{P}_m(R)$ for some $m > 0$ implies $\mathcal{F}_1(\bar{R}) \subseteq \mathcal{P}_{m-1}(\bar{R})$.*

Proof. We start observing that if R is a subperfect ring, then a module H satisfies $\operatorname{Tor}_1^R(H, Y) = 0$ for all torsion-free Y if and only if $H \in \mathcal{F}_1(R)$ (see [Fuchs 2019, Theorem 4.1]); here Y torsion-free means that $\operatorname{Tor}_1^R(R/tR, Y) = 0$ for all $t \in R^\times$. For any commutative ring R, $\operatorname{Tor}_1^R(X, Y) = 0$ for all torsion-free Y implies that $X \in \mathcal{F}_1(R)$ (but not necessarily conversely).

Recall [Cartan and Eilenberg 1956, Chapter VI, Proposition 4.1.1] which states that if an R-module Y satisfies $\operatorname{Tor}_k^R(\bar{R}, Y) = 0$ for all $k > 0$, then

$$\operatorname{Tor}_m^R(\bar{N}, Y) \cong \operatorname{Tor}_m^{\bar{R}}(\bar{N}, Y/rY) \tag{3}$$

holds for all $m > 0$ and for all $\bar{R}$-modules $\bar{N}$. The hypothesis holds if Y is a torsion-free R-module: it holds for $k = 1$ by definition and for $k > 1$ in view of $\text{p.d.}_R \bar{R} = 1$.

First, let $s \in R$ be a divisor of r, and choose $\bar{N} \cong R/sR$. Then the left-hand side Tor vanishes for all torsion-free Y and for $m = 1$, so it follows that Y/rY is a torsion-free $\bar{R}$-module.

(i) Assuming $\bar{R}$ is subperfect, let $\bar{N} \in \mathcal{F}_1(\bar{R})$ and Y a torsion-free R-module. The right hand side of (3) vanishes for $m = 1$, so we can conclude that $\operatorname{Tor}_1^R(\bar{N}, Y) = 0$. This equality holds for all torsion-free R-modules Y, whence we obtain $\bar{N} \in \mathcal{F}_1(R)$.

(ii) Assuming both R and $\bar{R}$ are subperfect, let again $\bar{N} \in \mathcal{F}_1(\bar{R})$. Part (i) implies that $\bar{N} \in \mathcal{F}_1(R)$, so $\bar{N} \in \mathcal{P}_m(R)$ by hypothesis. From a well-known Kaplansky

formula for projective dimensions [Kaplansky 1970, Proposition 172] we obtain that $\overline{N} \in \mathcal{P}_{m-1}(\overline{R})$, as claimed. □

Theorem 7.5. *If R is an n-subperfect ring, then* $\operatorname{Fdim}(R) \geq n$.

Proof. According to [Jensen 1972, Proposition 5.6], for any ring R, $\mathcal{F}_0(R) \subseteq \mathcal{P}_{m-1}(R)$ if $m = \operatorname{Fdim}(R) \geq 1$. Hence we have $\mathcal{F}_1(R) \subseteq \mathcal{P}_m(R)$. On the other hand, if R is n-subperfect, then Lemma 7.4 is applicable. Thus if $\mathcal{F}_1(R) \subseteq \mathcal{P}_k(R)$ holds for some k, then we have $\mathcal{F}_1(\overline{R}) \subseteq \mathcal{P}_{k-1}(\overline{R})$, and since R is n-subperfect, we can continue with $\overline{R}$ in the role of R, etc. If $k < n$, then this process would lead us down to $\mathcal{P}_0$, reaching a contradiction that over a subperfect ring of Krull dimension > 0 modules of weak dimension ≤ 1 are projective. Consequently, the inclusion $\mathcal{F}_1(R) \subseteq \mathcal{P}_k(R)$ can hold only if $k \geq n$. □

That we can have strict inequality in the preceding theorem is demonstrated by the following example. Let S denote an almost perfect (i.e., 1-subperfect) domain; it has finitistic dimension 1. If R is defined as in Example 8.1 as $S \oplus D$ with $D \neq 0$ a torsion-free divisible S-module, then $\operatorname{p.d.}_R R/D$ is finite and > 1 (D is flat, but not projective, so $\operatorname{p.d.}_R D = 1$), whence $\operatorname{Fdim}(R) \geq 2$.

The following result shows that in Theorem 7.5 equality may occur for non-Cohen–Macaulay rings as well.

Lemma 7.6. (i) *Let R be any ring, and $S = R[X]$ the polynomial ring over R. Then*

$$\mathcal{F}_1(R) \subseteq \mathcal{P}_n(R) \quad \textit{if and only if} \quad \mathcal{F}_1(S) \subseteq \mathcal{P}_{n+1}(S).$$

(ii) *If R is a perfect ring, then for the polynomial ring $S = R[X_1, \ldots, X_n]$* (which is n-subperfect by Theorem 6.2) *we have*

$$\mathcal{F}_1(S) \subseteq \mathcal{P}_n(S), \quad \textit{but} \quad \mathcal{F}_1(S) \nsubseteq \mathcal{P}_{n-1}(S).$$

Proof. (i) To verify necessity, assume M is a module in $\mathcal{F}_1(S)$. It is easy to see that then $M \in \mathcal{F}_1(R)$ as well, thus $M \in \mathcal{P}_n(R)$ follows by hypothesis. Hence tensoring over R with $R[X]$, we obtain $M[X] \in \mathcal{P}_n(S)$. It remains to refer to the exact sequence $0 \to M[X] \to M[X] \to M \to 0$ of S-modules to conclude that $M \in \mathcal{P}_{n+1}(S)$.

Conversely, working toward contradiction, suppose there are an $F \in \mathcal{F}_1(R)$ and an $H \in \text{Mod-}R$ such that $\operatorname{Ext}_R^{n+2}(F, H) \neq 0$. Then also $\operatorname{Ext}_R^{n+2}(F, H[X]) \neq 0$. Since $\operatorname{Tor}_k^R(F, S) = 0$ for all $k > 0$, we have an isomorphism (see [Cartan and Eilenberg 1956, Chapter VI, Proposition 4.1.3])

$$\operatorname{Ext}_S^{n+2}(F \otimes_R S, H[X]) \cong \operatorname{Ext}_R^{n+2}(F, H[X]) \neq 0.$$

Since $F \otimes_R S \in \mathcal{F}_1(S)$, this is in contradiction to $\mathcal{F}_1(S) \subseteq \mathcal{P}_{n+1}(S)$.

(ii) Noticing that $\mathcal{F}_1(R) = \mathcal{P}_0(R)$ if R is perfect, the claim follows by a simple calculation from (i). □

8. Examples

Our final section is devoted to various examples of n-subperfect rings. In the first examples we use n-subperfect domains to construct n-subperfect rings with nontrivial nilradicals. (For examples of nonnoetherian n-subperfect domains, we refer to Example 8.12 and Theorem 8.13 below.)

Example 8.1. Let S denote an n-subperfect domain ($n \geq 1$) with field of quotients H. Let D be a torsion-free divisible S-module. Define the ring R as the *idealization* of D, i.e., $R = S \oplus D$ additively, and multiplication in R is given by the rule

$$(s_1, d_1)(s_2, d_2) = (s_1 s_2, s_1 d_2 + s_2 d_1) \quad (s_i \in S,\ d_i \in D).$$

It is clear that $Q = (H, D)$ is the ring of quotients of R, and $N = (0, D)$ is the nilradical (nilpotent of exponent 2) of both R and Q. We claim that R is an n-subperfect ring.

First we observe that an element $r = (s, d) \in R$ is a zero-divisor if and only if $s = 0$; this is easily seen by direct calculation using the torsion-freeness of D. Hence criterion (iii) in Lemma 2.2 guarantees that R is a subperfect ring. Furthermore, for any $r = (s, d)$, we have $rR = (sS, D)$ (the divisibility of D is relevant). Therefore, we have an isomorphism $R/rR \cong S/sS$ for every regular $r \in R$ (i.e., for every nonzero $s \in S$). Hence we conclude that R/rR is $(n-1)$-subperfect for every regular r (Corollary 3.3). By the same Corollary, we obtain the desired conclusion for R.

Example 8.2. As before choose an n-subperfect ($n \geq 1$) integral domain S. Let A be any commutative S-algebra that is torsion-free and divisible as an S-module, and B a torsion-free divisible S-module containing A. Our ring R is now the ring of upper 3×3-triangular matrices of the form

$$\alpha = \begin{bmatrix} s & a & b \\ 0 & s & a \\ 0 & 0 & s \end{bmatrix} \quad (s \in S,\ a \in A,\ b \in B).$$

It is straightforward to check that $\alpha \in R$ is a zero-divisor if and only if $s = 0$, and that the principal ideal αR equals sR whenever $s \neq 0$. Fix any regular $\alpha_0 \in R$ (i.e., $0 \neq s_0 \in S$ in the diagonal), and consider the homomorphism $\phi : R \to S/s_0S$ given by $\alpha \mapsto s + s_0S$ ($\alpha \in R$). Then $\operatorname{Ker}\phi = \alpha_0 R = s_0 R$ leads to the isomorphism $R/\alpha_0 R \cong S/s_0S$ showing that $R/\alpha_0 R$ is an $(n-1)$-subperfect ring for every regular $\alpha_0 \in R$. To complete the proof that R is n-subperfect, it remains only to show that R is subperfect. By Lemma 2.2(iii) it suffices to observe that the nilradical N of R

is nilpotent of degree 3, and every regular coset mod N consists of regular elements of R.

In order to obtain more general examples of similar kind, in the preceding examples we can choose S as a finite direct sum of n-subperfect domains.

Let R be a perfect ring, and let $S = R[X_1, \ldots, X_n]$. By Corollary 5.2 and Theorem 6.2, the ring of invariants S^G of S is n-subperfect for each finite group G acting on S whose order is a unit in R. As in the classical case in which R is a field, more examples of n-subperfect rings can be obtained from S via Veronese subrings: a *Veronese subring* T of S is an R-subalgebra of S generated by all monomials of degree d for some fixed $d > 0$.

Theorem 8.3. *Let R be a ring, and $S = R[X_1, \ldots, X_n]$ a polynomial ring over R. A Veronese subring of S is n-subperfect if and only if R is perfect.*

Proof. Let T be a Veronese subring of S generated by the monomials of degree d. Then T is an R-direct summand of S, and S is integral over T. If R is perfect, then T is n-subperfect by Theorems 5.1 and 6.2. Conversely, suppose T is n-subperfect. Then $X_1^d, \ldots, X_n^d$ is a maximal regular sequence of T, so $T/(X_1^d, \ldots, X_n^d)R$ is a perfect ring. As a homomorphic image of this ring, R is perfect. □

Remark 8.4. Asgharzadeh, Dorreh and Tousi [Asgharzadeh et al. 2017] study Cohen–Macaulay properties for Veronese, determinantal, and Grassmannian rings in the context of polynomial rings in infinitely many variables.

Theorem 6.2 shows that if R is perfect, then the ring $R[X_1, \ldots, X_n]$ is n-subperfect. As the next example demonstrates, it need not be the case that for a k-subperfect ring R, $R[X_1, \ldots, X_n]$ is $(n+k)$-subperfect.

Example 8.5. Let F be a field, X, Y indeterminates, and $K = F(X)$. Then the ring $R = F + YK[\![Y]\!]$ is an almost perfect domain [Bazzoni and Salce 2003, Example 3.2]. The valuative dimension of R, that is, the maximum of the Krull dimensions of the valuation rings of $Q(R)$ that contain R, is 2. Thus $\dim R[X_1, X_2] = 4$ by [Arnold 1969, Theorem 6]. Although R is 1-subperfect, $R[X_1, X_2]$ is not 3-subperfect, since by Corollary 3.6 the Krull dimension of a 3-subperfect ring is 3.

Example 8.6. *An n-subperfect ($n \geq 1$) Prüfer domain is a Dedekind domain* (hence 1-subperfect). First of all, a Prüfer domain R cannot have a regular sequence of length greater than 1. Indeed, if x, y is a regular sequence in R, then $xR \cap yR = xyR$. If M is a maximal ideal containing x and y, then since R_M is a valuation domain, this implies $xR_M = xyR_M$ or $yR_M = xyR_M$, contradicting that neither x nor y is a unit in R_M. Consequently, an n-subperfect Prüfer domain is an almost perfect domain. But for modules over such domains, w.d. ≤ 1 implies p.d. ≤ 1 (see [Fuchs and Salce 2018, Theorem 6.1]), thus any n-subperfect Prüfer domain — if not a field — must be a Dedekind domain. Dedekind domains are trivially 1-subperfect.

Our next source of examples involves the idealization of a module, as defined in Example 8.1. For an R-module N, we denote by $R \star N$ the idealization of N. It is well known that if R is a Cohen–Macaulay ring and N is a maximal Cohen–Macaulay module, then $R \star N$ is a Cohen–Macaulay ring. In Corollary 8.8, we prove the analogue of this statement for n-subperfect rings. This follows from a more general lifting property of n-subperfectness:

Theorem 8.7. *Let I be an ideal of the ring R such that $I^2 = 0$ and R/I is n-subperfect for some $n \geq 0$. If every (R/I)-regular sequence in R is also I-regular, then R is n-subperfect.*

Proof. First we show that R is subperfect. If N is the nilradical of R, then N/I is the nilradical of the n-subperfect ring R/I, hence T-nilpotent. Therefore, N as an extension of the nilpotent I by the T-nilpotent N/I is T-nilpotent. Suppose $r + N$ ($r \in R$) is a regular element in R/N; then $r + N/I$ is regular in $(R/I)/(N/I)$, so Lemma 2.2(iii) shows that $r + I$ is regular in R/I. Since r is (R/I)-regular, r is I-regular by assumption. If r is both (R/I)-regular and I-regular, then it is regular in R. From Lemma 2.2(iii) we conclude that R is subperfect.

We claim next that each $r \in R^\times$ is (R/I)-regular. Since R/I is subperfect, there are finitely many prime ideals $P_1, \ldots, P_m$ of R that are minimal over I and whose images in R/I contain every zero-divisor in R/I. Since I is in the nilradical of R, these primes are also the minimal prime ideals of R. If $r \in R^\times$, then $r \notin P_1 \cup \cdots \cup P_m$, so the image of r in R/I is not a zero-divisor. This shows that the regular elements of R are (R/I)-regular.

We prove now using induction that R is n-subperfect. If $n = 0$, then R/I is perfect and hence zero-dimensional. Since $I^2 = 0$, R is zero-dimensional. We have established that R is subperfect, so from $R = Q(R)$ we conclude that R is perfect.

Now let $n > 0$, and suppose the theorem has been proved for all $k < n$. We have already shown that R is subperfect. We claim that $A := R/rR$ is $(n-1)$-subperfect for every $r \in R^\times$. By the induction hypothesis, it suffices to show

(i) $(IA)^2 = 0$,

(ii) A/IA is $(n-1)$-subperfect, and

(iii) every A/IA-regular sequence in A is IA-regular.

It is clear that $(IA)^2 = 0$. To verify (ii), we use the fact already established that if $r \in R^\times$, then $r + I$ is regular in R/I. Since R/I is n-subperfect, Proposition 3.2 implies $R/(rR + I)$ is $(n-1)$-subperfect. In view of the isomorphism $A/IA \cong R/(rR + I)$, statement (ii) follows.

To verify (iii), suppose $a_1, \ldots, a_t$ is an A/IA-regular sequence in A. If we write $a_i = r_i + rR$, then $r_1, \ldots, r_t$ is an A/IA-regular sequence in R. Since $r \in R^\times$ and $A/IA \cong R/(rR + I)$, we have that $r, r_1, \ldots, r_n$ is an R/I-regular sequence. By

assumption, $r, r_1, \ldots, r_t$ is also an I-regular sequence, so $r_1, \ldots, r_t$ is an I/rI-regular sequence. As established, every regular element of R is a regular element in R/I. Thus $I \cap rR = rI$, and it follows that $IA = (I+rR)/rR \cong I/(I \cap rR) = I/rI$. Since $r_1, \ldots, r_t$ is an (I/rI)-regular sequence in R, we conclude that $a_1, \ldots, a_t$ is an IA-regular sequence in A. Thus every A/IA-regular sequence in A is IA-regular.

Having verified (i), (ii) and (iii), we conclude from the induction hypothesis that $A = R/rR$ is $(n-1)$-subperfect. Since R is subperfect and R/rR is $(n-1)$-subperfect for each $r \in R^\times$, Corollary 3.3 implies R is n-subperfect. □

Corollary 8.8. *Let R be an n-subperfect ring, and let N be an R-module such that every regular sequence in R extends to a regular sequence on N. Then $R \star N$ is an n-subperfect ring.* □

Example 8.9. Corollary 8.8 implies that if R is a local Cohen–Macaulay ring, and if N is a balanced big Cohen–Macaulay R-module, then $R \star N$ is n-subperfect for $n = \dim R$. Choosing N to be an infinite rank free R-module, we obtain a nonnoetherian n-subperfect ring $R \star N$.

More interesting choices are possible for N. For example if R is an excellent local Cohen–Macaulay domain of positive characteristic and positive dimension, and R^+ is the integral closure of R in the algebraic closure of the quotient field of R, then $R \star R^+$ is a nonnoetherian n-subperfect ring, since R^+ is a balanced big Cohen–Macaulay module that is not finitely generated [Hochster and Huneke 1992, Theorem 1.1].

Example 8.10. Let R be an n-subperfect ring and $\{X_i : i \in I\}$ a collection of indeterminates for R. Let

$$S = R[X_i : i \in I]/(X_i : i \in I)^2.$$

The ideal $N = (X_i : i \in I)/(X_i : i \in I)^2$ of S is nilpotent of index 2 and is a free R-module with basis the images of the X_i in N. As $S \cong R \star N$, the ring S is a special case of the construction in Example 8.9; therefore, S is n-subperfect. If the index set I is infinite, then S is not noetherian.

So far, our nonnoetherian examples, at least for $n > 1$, have involved n-subperfect rings with zero-divisors. Our next source of examples produces nonnoetherian n-subperfect domains, albeit in a nontransparent way.

Theorem 8.11. *Let S be a local Cohen–Macaulay domain such that $Q(S)$ is separably generated, and has positive characteristic and uncountable transcendence degree over its prime subfield. If $n := \dim S \geq 1$, then there exists a nonnoetherian n-subperfect subring R of S such that $Q(R) = Q(S)$ and S is integral over R.*

Proof. Let N be a free S-module of infinite rank. Applying [Olberding 2012, Theorem 3.5] to S and N, we obtain a subring R of S such that R is "strongly

twisted by N." We omit the definition of this notion here, but we use the fact that by [Olberding 2012, Theorems 4.1 and 4.6] this implies

(i) there is a subring A of R such that S/A is a torsion-free divisible A-module and $I \cap A \neq 0$ for each ideal I of S;

(ii) R has the same quotient field as S and S is an integral extension of R; and

(iii) there is a faithfully flat ring embedding $f : R \to S \star N$ such that for each $0 \neq a \in A$, the induced map $f_a : R/aR \to (S \star N)/a(S \star N)$ is an isomorphism.

We show that the ring $R/(x_1, \dots, x_t)R$ is subperfect for each nonempty regular sequence $x_1, \dots, x_t$ in R. Since f is faithfully flat, $f(x_1), \dots, f(x_t)$ is a regular sequence in $T := S \star N$. By Corollary 8.8, T is an n-subperfect ring. Thus $f(x_1), \dots, f(x_t)$ is a subperfect sequence in T. Since for each $0 \neq a \in A$, the map f_a is an isomorphism, we have $T = f(R) + f(a)T$. By (i) and (ii), the fact that S/R is a torsion R-module implies there is $0 \neq a \in (x_1, \dots, x_t)R \cap A$. Hence

$$T = f(R) + (f(x_1), \dots, f(x_t))T.$$

Moreover, since f is faithfully flat, we have

$$(f(x_1), \dots, f(x_t))T \cap f(R) = (f(x_1), \dots, f(x_t))f(R).$$

Therefore,

$$\begin{aligned} T/(f(x_1), \dots, f(x_t))T &= \big(f(R) + (f(x_1), \dots, f(x_t))T\big)/(f(x_1), \dots, f(x_t))T \\ &\cong f(R)/\big((f(x_1), \dots, f(x_t))T \cap f(R)\big) \\ &= f(R)/(f(x_1), \dots, f(x_t))f(R) \\ &\cong R/(x_1, \dots, x_t)R. \end{aligned}$$

Consequently, since $f(x_1), \dots, f(x_t)$ is a subperfect sequence in T, it follows that $x_1, \dots, x_t$ is a subperfect sequence in R. This proves that every regular sequence in R is subperfect.

Finally, since S is integral over R and S is local, R is also local and has the same Krull dimension as S. By Corollary 3.6, $n = \dim S = \dim R$. Taking into account that every regular sequence in R is subperfect, Corollary 3.9 implies that R is n-subperfect. By [Olberding 2012, Theorem 5.2], the fact that N is a free S-module of infinite rank implies R is not noetherian. □

Example 8.12. Let p be a prime number, and let $\mathbb{F}_p$ denote the field with p elements. Suppose k is a purely transcendental extension of $\mathbb{F}_p$ with uncountable transcendence degree. Then $S = k[X_1, \dots, X_n]_{(X_1, \dots, X_n)}$ is a local n-subperfect domain (in fact, a Cohen–Macaulay ring) meeting the requirements of Theorem 8.11. Thus S contains a nonnoetherian n-subperfect subring R having the same quotient field as S.

Our final source of examples involves local Cohen–Macaulay rings that have a coefficient field. The next theorem shows that restriction to a smaller coefficient field can produce examples of nonnoetherian n-subperfect rings.

Theorem 8.13. *Let S be a local Cohen–Macaulay ring containing a field F such that $S = F + M$, where M is the maximal ideal of S. For each subfield k of F, the local ring $R = k + M$ is n-subperfect for $n = \dim S$. The ring R is noetherian if and only if F/k is a finite extension.*

Proof. Evidently, R is a local ring with maximal ideal M. It is clear that every prime ideal of S is a prime ideal of R. To verify the converse, let P be a nonmaximal prime ideal of R. To show that P is in fact an ideal of S, let $s \in S$. Then $sP \subseteq sM \subseteq R$, and also, $(sP)M = P(sM) \subseteq P$ because $sM \subseteq R$. Since $M \not\subseteq P$, we conclude that $sP \subseteq P$, which proves that P is an ideal of S. To see that P is prime in S, let $x, y \in S$ with $xy \in P$. If one of x or y is a unit in S, then the other is in P. Otherwise, if neither x nor y are units, then necessarily $x, y \in M \subseteq R$, and since P is a prime ideal of R, one of x, y is in P. Thus P is a prime ideal of S, and this shows that the prime ideals of R are precisely those of S.

We show now that R is n-subperfect, where $n = \dim S$. By [Fontana et al. 1997, Lemma 1.1.4, p. 5], $Q(R) = Q(S)$, so R is a subperfect ring, since the total quotient ring $Q(S)$ of the Cohen–Macaulay ring S is artinian. Let $x_1, \ldots, x_t$ be a regular sequence in R, and $I = (x_1, \ldots, x_t)R$. We claim that R/I is a subperfect ring. The height of I in R is at least t, and since R and S share the same prime ideals, the height of IS is also at least t. Krull's height theorem implies then that the height of the t-generated ideal IS is t. Since S is a Cohen–Macaulay ring, the ideal IS is unmixed. We use this to show next that $Q(R/I)$ is zero-dimensional.

To this end, we prove that every zero-divisor of R/I is contained in a minimal prime ideal of R/I. Let $x, y \in R$ such that $xy \in I$ and $y \notin I$. Suppose by way of contradiction that x is not contained in any minimal prime ideal of I. Since I and IS share the same minimal primes, the image of x in S/IS does not belong to any minimal prime ideal of S/IS. However, IS is unmixed, so necessarily $y \in IS$. Therefore, using the fact that $S = F + M$, we can write

$$y = \alpha_1 x_1 + \cdots + \alpha_t x_t + z \quad \text{for } \alpha_1, \ldots, \alpha_t \in F \text{ and } z \in (x_1, \ldots, x_t)M.$$

Similarly, since $xy \in I$ and $R = k + M$, we have

$$xy = \beta_1 x_1 + \cdots + \beta_t x_t + w \quad \text{for } \beta_1, \ldots, \beta_t \in k \text{ and } w \in (x_1, \ldots, x_t)M.$$

Let i be the largest index such that at least one of α_i, β_i is not 0. Using the preceding expressions for y and xy, we obtain

$$\beta_1 x_1 + \cdots + \beta_i x_i + w = \alpha_1 x x_1 + \cdots + \alpha_i x x_i + xz.$$

Therefore,

$$(\beta_i - \alpha_i x)x_i \in (x_1, \ldots, x_{i-1})R.$$

Since $\beta_i - \alpha_i x \in k + M = R$ and $x_1, \ldots, x_i$ is a regular sequence in R, we have $\beta_i - \alpha_i x \in (x_1, \ldots, x_{i-1})R$. The fact that x is a nonunit in R implies $\beta_i \in M$, so $\beta_i = 0$ and hence, by the choice of i, $\alpha_i \neq 0$. Since the prime ideals of S are the same as the prime ideals of R, $\sqrt{(x_1, \ldots, x_{i-1})R}$ is an ideal of S. Also, α_i is a unit in S and $\alpha_i x \in (x_1, \ldots, x_{i-1})R$, so $x \in \sqrt{(x_1, \ldots, x_{i-1})R} \subseteq \sqrt{I}$. However, x was chosen not to be contained in any minimal prime ideal of I. This contradiction implies that x must be in some minimal prime ideal of I, establishing that $Q(R/I)$ is a zero-dimensional ring. Since I and IS share the same minimal prime ideals, I has only finitely many minimal primes, so $Q(R/I)$ is also semilocal.

It remains to show that the nilradical of R/I is T-nilpotent, and to prove this, it suffices to show that some power of $\sqrt{I}$ is contained in I. Since $\sqrt{I}$ is a finitely generated ideal of the noetherian ring S and $\sqrt{I} = \sqrt{IM}$, with IM an ideal of S, there is $t > 0$ such that $(\sqrt{I})^t \subseteq IM \subseteq I$. Therefore, R/I is subperfect, which completes the proof that every regular sequence in R is subperfect. Since R and S share the same maximal ideal, Corollary 4.3 implies R is n-subperfect for $n = \dim S$. Finally, it is straightforward to check that R is noetherian if and only if F/k is a finite field extension; see [Fontana et al. 1997, Proposition 1.1.7, p. 7]. □

Example 8.14. Let $S = F[\![X_1, \ldots, X_n]\!]/I$, where F is a field, $X_1, \ldots, X_n$ are indeterminates for F, and I is an ideal such that S is Cohen–Macaulay. Theorem 8.13 implies that for each subfield k of F,

$$R = \left\{ f + I \in S : f \in F[\![X_1, \ldots, X_n]\!] \text{ and } f(0, \ldots, 0) \in k \right\}$$

is an n-subperfect ring.

Acknowledgment

We thank the referee for a careful reading of the paper and suggestions that improved the clarity of the arguments.

References

[Arapović 1983] M. Arapović, "Characterizations of the 0-dimensional rings", *Glas. Mat. Ser. III* **18(38)**:1 (1983), 39–46. MR Zbl

[Arnold 1969] J. T. Arnold, "On the dimension theory of overrings of an integral domain", *Trans. Amer. Math. Soc.* **138** (1969), 313–326. MR Zbl

[Asgharzadeh and Tousi 2009] M. Asgharzadeh and M. Tousi, "On the notion of Cohen–Macaulayness for non-Noetherian rings", *J. Algebra* **322**:7 (2009), 2297–2320. MR Zbl

[Asgharzadeh et al. 2017] M. Asgharzadeh, M. Dorreh, and M. Tousi, "Direct summands of infinite-dimensional polynomial rings", *J. Commut. Algebra* **9**:1 (2017), 1–19. MR Zbl

[Bass 1960] H. Bass, "Finitistic dimension and a homological generalization of semi-primary rings", *Trans. Amer. Math. Soc.* **95** (1960), 466–488. MR Zbl

[Bass 1963] H. Bass, "On the ubiquity of Gorenstein rings", *Math. Z.* **82** (1963), 8–28. MR Zbl

[Bazzoni and Salce 2003] S. Bazzoni and L. Salce, "Almost perfect domains", *Colloq. Math.* **95**:2 (2003), 285–301. MR Zbl

[Björk 1969] J.-E. Björk, "Rings satisfying a minimum condition on principal ideals", *J. Reine Angew. Math.* **236** (1969), 112–119. MR Zbl

[Bruns and Herzog 1998] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, revised ed., Cambridge Studies in Advanced Mathematics **39**, Cambridge University Press, 1998. Zbl

[Cartan and Eilenberg 1956] H. Cartan and S. Eilenberg, *Homological algebra*, Princeton University Press, 1956. MR Zbl

[Coleman and Enochs 1971] D. B. Coleman and E. E. Enochs, "Isomorphic polynomial rings", *Proc. Amer. Math. Soc.* **27** (1971), 247–252. MR Zbl

[Fontana et al. 1997] M. Fontana, J. A. Huckaba, and I. J. Papick, *Prüfer domains*, Monographs and Textbooks in Pure and Applied Mathematics **203**, Marcel Dekker, New York, 1997. MR Zbl

[Fuchs 2019] L. Fuchs, "Characterizing almost perfect rings by covers and envelopes", *J. Korean Math. Soc.* (online publication September 2019).

[Fuchs and Salce 2018] L. Fuchs and L. Salce, "Almost perfect commutative rings", *J. Pure Appl. Algebra* **222**:12 (2018), 4223–4238. MR Zbl

[Glaz 1994] S. Glaz, "Coherence, regularity and homological dimensions of commutative fixed rings", pp. 89–106 in *Commutative algebra* (Trieste, 1992), World Sci. Publ., River Edge, NJ, 1994. MR Zbl

[Goodearl and Warfield 1989] K. R. Goodearl and R. B. Warfield, Jr., *An introduction to noncommutative Noetherian rings*, London Mathematical Society Student Texts **16**, Cambridge University Press, 1989. MR Zbl

[Gupta 1970] R. N. Gupta, "Characterization of rings whose classical quotient rings are perfect rings", *Publ. Math. Debrecen* **17** (1970), 215–222. MR Zbl

[Hamilton 2004] T. D. Hamilton, "Weak Bourbaki unmixed rings: a step towards non-Noetherian Cohen–Macaulayness", *Rocky Mountain J. Math.* **34**:3 (2004), 963–977. MR Zbl

[Hamilton and Marley 2007] T. D. Hamilton and T. Marley, "Non-Noetherian Cohen–Macaulay rings", *J. Algebra* **307**:1 (2007), 343–360. MR Zbl

[Hochster and Eagon 1971] M. Hochster and J. A. Eagon, "Cohen–Macaulay rings, invariant theory, and the generic perfection of determinantal loci", *Amer. J. Math.* **93** (1971), 1020–1058. MR Zbl

[Hochster and Huneke 1992] M. Hochster and C. Huneke, "Infinite integral extensions and big Cohen–Macaulay algebras", *Ann. of Math.* (2) **135**:1 (1992), 53–89. MR Zbl

[Jensen 1972] C. U. Jensen, *Les foncteurs dérivés de* $\varprojlim$ *et leurs applications en théorie des modules*, Lecture Notes in Mathematics **254**, Springer, 1972. MR Zbl

[Kabele 1971] T. Kabele, "Regularity conditions in nonnoetherian rings", *Trans. Amer. Math. Soc.* **155** (1971), 363–374. MR Zbl

[Kaplansky 1970] I. Kaplansky, *Commutative rings*, Allyn and Bacon, Boston, 1970. MR Zbl

[Lam 1999] T. Y. Lam, *Lectures on modules and rings*, Graduate Texts in Mathematics **189**, Springer, New York, 1999. MR Zbl

[Lam 2001] T. Y. Lam, *A first course in noncommutative rings*, 2nd ed., Graduate Texts in Mathematics **131**, Springer, 2001. MR Zbl

[Lee 2006] S. B. Lee, "Weak-injective modules", *Comm. Algebra* **34**:1 (2006), 361–370. MR Zbl

[Lipman 1965] J. Lipman, "Some properties of localization and normalization", *Proc. Amer. Math. Soc.* **16** (1965), 1120–1122. MR Zbl

[Matsumura 1986] H. Matsumura, *Commutative ring theory*, Cambridge Studies in Advanced Mathematics **8**, Cambridge University Press, 1986. MR Zbl

[Ohm and Pendleton 1968] J. Ohm and R. L. Pendleton, "Rings with noetherian spectrum", *Duke Math. J.* **35** (1968), 631–639. MR Zbl

[Olberding 2012] B. Olberding, "A counterpart to Nagata idealization", *J. Algebra* **365** (2012), 199–221. MR Zbl

[Olberding 2019] B. Olberding, "Height theorems and unmixedness for polynomial rings over zero-dimensional rings", preprint, 2019.

[Sandomierski 1973] F. L. Sandomierski, "Homological dimension under change of rings", *Math. Z.* **130** (1973), 55–65. MR Zbl

[Small 1966] L. W. Small, "Orders in Artinian rings", *J. Algebra* **4** (1966), 13–41. MR Zbl

Received January 6, 2018. Revised April 18, 2019.

LÁSZLÓ FUCHS
DEPARTMENT OF MATHEMATICS
TULANE UNIVERSITY
NEW ORLEANS, LA
UNITED STATES

fuchs@tulane.edu

BRUCE OLBERDING
DEPARTMENT OF MATHEMATICAL SCIENCES
NEW MEXICO STATE UNIVERSITY
PO BOX 30001, MSC MB
LAS CRUCES, NM
UNITED STATES

olberdin@nmsu.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.165

ORDINARY POINTS MOD p OF $\mathrm{GL}_n(\mathbb{R})$-LOCALLY SYMMETRIC SPACES

MARK GORESKY AND YUNG SHENG TAI

Locally symmetric spaces for $\mathrm{GL}_n(\mathbb{R})$ parametrize polarized complex abelian varieties with real structure (antiholomorphic involution). We introduce a mod p analog. We define an "antiholomorphic" involution (or "real structure") on an ordinary abelian variety (defined over a finite field k) to be an involution of the associated Deligne module (T, F, V) that exchanges F (the Frobenius) with V (the Verschiebung). The definition extends to include principal polarizations and level structures. We show there are finitely many isomorphism classes of such objects in each dimension, and give a formula for this number that resembles the Kottwitz "counting formula" (for the number of principally polarized abelian varieties over k), but the symplectic group in the Kottwitz formula has been replaced by the general linear group.

1. Introduction

1.1. Let $N \geq 3$ be an integer and let $\Gamma_N \subset \mathrm{Sp}_{2n}(\mathbb{Z})$ be the principal level N subgroup consisting of elements that are congruent to the identity modulo N. The locally symmetric space $Y = \Gamma_N \backslash \mathrm{Sp}_{2n}(\mathbb{R})/U(n)$ may be viewed as the set of complex points of the moduli space $\mathcal{A}_{n,[N]}$ of principally polarized complex n-dimensional abelian varieties with level N structure. It admits the structure of a complex algebraic variety and it has an incarnation "modulo p", namely, the moduli space $\mathcal{A}_{n,[N]}(k)$ of principally polarized abelian varieties (of dimension n with level N structure) over a finite field $k = \mathbb{F}_q$ of characteristic p. The number of points in $\mathcal{A}_{n,[N]}(k)$ was computed by R. Kottwitz [1990; 1992], proving a reformulation of the conjecture of Langlands and Rapoport [1987], following earlier work on this question by J. Milne, W. Waterhouse, R. Langlands, M. Rapoport and others.

For $n \geq 3$ the locally symmetric space $X = \mathrm{GL}_n(\mathbb{Z}) \backslash \mathrm{GL}_n(\mathbb{R})/\mathrm{O}(n)$ does not have a complex structure. Nevertheless in many ways this space behaves something like an algebraic variety, perhaps most spectacularly illustrated by the success (see [Harris et al. 2016; Harris and Taylor 2001; Patrikis and Taylor 2015; Barnet-Lamb et al. 2014; Taylor 2008; Scholze 2015]) in associating Galois representations to

MSC2010: 11F99, 11G25, 14G35, 14K10.
Keywords: ordinary abelian variety, locally symmetric space, Kottwitz formula.

modular forms on X. This leads to the search for other ways in which the locally symmetric space X behaves like the algebraic variety Y. Is it possible to make sense of the points of X "modulo p", and to provide a concrete description and count for the points of X over the finite field $\mathbb{F}_q$?

With appropriate level structures, finitely many copies of the space X sit inside Y in a natural way. In [Goresky and Tai 2003a], we showed that the (principally polarized) abelian varieties corresponding to points $x \in X$ are precisely those which admit a real structure, that is, an antiholomorphic involution. Therefore one might hope to identify the finite field analog of X as a parameter space for principally polarized abelian varieties over $\mathbb{F}_q$ equipped with an "antiholomorphic involution", whatever that means.

1.2. A hint is provided by the theory of complex multiplication. If a simple CM abelian variety A has good reduction to a variety $\bar{A}$ over $\mathbb{F}_q$ then the Frobenius morphism F has a lift to an element $\pi \in \mathrm{End}_{\mathbb{Q}}(A)$. Complex conjugation takes π to $\bar{\pi} = q\pi^{-1}$ (since π is a Weil q-number) which is a lift of the Verschiebung V. Therefore if "complex conjugation" is to make sense on $\bar{A}$ it must switch F and V. Is it possible to enlarge the collection of morphisms for abelian varieties over $\mathbb{F}_q$ so as to allow for generalized morphisms that switch the Frobenius with the Verschiebung?

1.3. In this paper we show how to make sense of these notions for *ordinary* abelian varieties over $\mathbb{F}_q$ using P. Deligne's linear algebra description [1969] of the category of ordinary abelian varieties as equivalent to the category of Deligne modules (T, F). We define a real structure on (T, F) to be an involution $\tau : T \to T$ that switches F and $V = qF^{-1}$. This simple, almost trivial definition leads to a wealth of interesting structures. The definition extends naturally to include polarizations (using Howe's theorem [1995]) and level structures so we obtain a category of "real" polarized Deligne modules. We show there are finitely many isomorphism classes of real Deligne modules (T, F, τ) (with principal polarization and level structure) over $\mathbb{F}_q$, and we are able to count them. For $n = 1$ in Section 7.4, we find, asymptotically $C(p)q^{1/2}\log q$ objects (for q an odd power of p). For general n we show that the method of Kottwitz [1990; 1992] may be modified to give a formula, involving adèlic orbital integrals at p and away from p, that closely resembles the finite adèlic part of the (relative) trace formula.

1.4. Conceptually, the general formula (Section 10.5) may be described as follows. By appropriate choice of coordinates it turns out that the Frobenius morphism F for an ordinary abelian variety with real structure (that is, for a polarized Deligne module with involution (T, F, τ)) may be expressed as a semisimple element

$$\gamma_0 = \begin{pmatrix} A & B \\ C & {}^tA \end{pmatrix} \in \mathrm{GSp}_{2n}(\mathbb{Q})$$

such that B, C are symmetric and A is self-adjoint with respect to the inner product

defined by C, and is totally real[1] with eigenvalues of absolute value $< \sqrt{q}$. It turns out that the blocks A, B, C have elegant interpretations: the $\mathrm{GL}_n(\mathbb{Q})$-conjugacy class of A determines the $\overline{\mathbb{Q}}$-isogeny class of (T, F, τ), reflecting the equivalence of conjugacy and stable conjugacy for GL_n. Moreover, the congruence class of C determines the $\mathbb{Q}$-isogeny class of (T, F, τ) within its $\overline{\mathbb{Q}}$-isogeny class (and B is uniquely determined by A, C). The number of isomorphism classes within a $\mathbb{Q}$-isogeny class is given by an orbital integral.

1.5. Our formula differs from that of [Kottwitz 1990] in that the contribution "at p" is an ordinary orbital integral as opposed to the twisted orbital integral that arises in [Kottwitz 1990]. Kottwitz uses a special case of the fundamental lemma to express the twisted integral in terms of (stable) ordinary integrals. In our case we do the reverse: in [Goresky and Tai 2019] (which is not restricted to the "ordinary" case), by comparing $\mathbb{Z}_p$-lattices with lattices over the Witt vectors, we show that the contribution "at p" to our formula can also be expressed as a (single) twisted orbital integral, which in turn can be interpreted as counting Dieudonné modules with antiholomorphic involution.

1.6. Because we restrict to the "ordinary" case, most of the techniques of this paper involve little more than linear algebra. In some sections, for completeness we have provided proofs of results that are known to experts. As a byproduct, we obtain an elementary re-proof of the "ordinary" part of the Kottwitz formula (see Theorem 10.2). It is simpler than the general formula because it does not require the Kottwitz invariant $\alpha(\gamma_0; \gamma\delta)$, and does not involve a twisted orbital integral, but we include it because it provides a useful comparison with the formula in Section 10.5 in the "real" case, in which the symplectic group has been replaced by the general linear group.

1.7. ***Notation.*** If E is an algebraic number field, we use $\mathcal{O}_E$ to denote its full ring of integers. Throughout this paper we fix a finite field $k = \mathbb{F}_q$ of characteristic $p > 0$. If R is an integral domain and $n \geq 1$, the *standard symplectic form* $\omega_0 : R^{2n} \times R^{2n} \to R$ is the bilinear map whose matrix is $\left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$. The general symplectic group, which we denote by $G(R) = \mathrm{GSp}_{2n}(R)$, consists of elements $A \in \mathrm{GL}_{2n}(R)$ such that $\omega_0(Ax, Ay) = \lambda\omega_0(x, y)$ for some $\lambda \in R^\times$ in which case λ is a character, called the *multiplier*. The *standard involution* (see Appendix B) on R^{2n} is the map $\tau_0(x, y) = (-x, y)$. If $g \in G(R)$, we set $\tilde{g} = \tau_0 g \tau_0^{-1}$. The subgroup fixed under this involution is denoted $\mathrm{GL}_n^*(R)$ (see Section 5.4). The finite adèles of $\mathbb{Q}$ are denoted $\mathbb{A}_f$. Let

$$\mathbb{A}_f^p = \prod\nolimits'_{v \neq p, \infty} \mathbb{Q}_p$$

denote the adèles away from p, let $\widehat{\mathbb{Z}}^p = \prod_{v \neq p, \infty} \mathbb{Z}_v$ so that $\widehat{\mathbb{Z}} = \mathbb{Z}_p . \widehat{\mathbb{Z}}^p$. Let

[1] Meaning that its eigenvalues are totally real algebraic integers.

$K_N \subset \mathrm{GSp}_{2n}(\mathbb{Z})$ and $K_N^0 \subset \mathrm{Sp}_{2n}(\mathbb{Z})$ denote the principal congruence subgroups of level N and similarly

$$\begin{aligned} \widehat{K}_N^0 &= \ker(\mathrm{Sp}_{2n}(\widehat{\mathbb{Z}}) \to \mathrm{Sp}_{2n}(\mathbb{Z}/M\mathbb{Z})), \\ \widehat{K}_N &= \ker(\mathrm{GSp}_{2n}(\widehat{\mathbb{Z}}) \to \mathrm{GSp}_{2n}(\mathbb{Z}/N\mathbb{Z})) = \widehat{K}_N^p K_p, \end{aligned} \tag{1.7.1}$$

where $\widehat{K}_N^p = \mathrm{GSp}_{2n}(\widehat{\mathbb{Z}}^p) \cap \widehat{K}_N$ and $K_p = \mathrm{GSp}_{2n}(\mathbb{Z}_p)$. If S is a commutative ring with 1 and $\mathcal{C}$ is a $\mathbb{Z}$-linear abelian category, the associated category *up to S-isogeny* ([Deligne 1973; Kottwitz 1990]) is the category with the same objects but with morphisms

$$\mathrm{Hom}_S(A, B) = \mathrm{Hom}_{\mathcal{C}}(A, B) \otimes_{\mathbb{Z}} S.$$

An S-isogeny is an isomorphism in this category, i.e, an invertible element in this set.

2. The complex case

2.1. We briefly recall several aspects of the theory of moduli of real abelian varieties, which serve as a partial motivation for the results in this paper. Recall that a real structure on a complex abelian variety A is an antiholomorphic involution of A. It has been observed [Silhol 1982; Seppälä and Silhol 1989; Comessatti 1926; Shimura 1975; Milne and Shih 1981; Adler 1979; Gross and Harris 1981; Goresky and Tai 2003b] that principally polarized abelian varieties (of dimension n) with real structure correspond to "real points" of the coarse moduli space

$$Y = \mathrm{Sp}_{2n}(\mathbb{Z})\backslash\mathfrak{h}_n$$

of all principally polarized abelian varieties, where $\mathfrak{h}_n$ is the Siegel upper halfspace. On this variety, complex conjugation is induced from the mapping on $\mathfrak{h}_n$ that is given by $Z \mapsto \tilde{Z} = -\bar{Z}$ which is in turn induced from the "standard involution" τ_0.

However, a given principally polarized abelian variety A may admit several nonisomorphic real structures ([Silhol 1982]). Thus, the coarse moduli space of principally polarized abelian varieties with real structure is not a subset of Y but rather, it maps to Y by a finite mapping. This multiplicity may be removed by replacing Y with the moduli space of principally polarized abelian varieties with a sufficiently high level structure. More generally let $K_f \subset \mathrm{Sp}_{2n}(\mathbb{A}_f)$ be a compact open subgroup of the finite adèlic points of Sp_{2n} that is preserved by the involution τ_0 and is sufficiently small that $K_f \cap \mathrm{Sp}_{2n}(\mathbb{Q})$ is torsion free. (We use Sp rather than GSp for expository purposes because the argument for GSp is similar but slightly messier.) As in [Rohlfs 1978], the fixed points of the involution τ_0 on double coset space

$$Y = \mathrm{Sp}_{2n}(\mathbb{Q})\backslash\mathrm{Sp}_{2n}(\mathbb{A})/K_f U(n)$$

are classified by classes[2] in the nonabelian cohomology $H^1(\langle\tau_0\rangle, K_f)$.

[2] If $x \in \mathrm{Sp}_{2n}(\mathbb{A})$ maps to a fixed point in Y there exist $\gamma \in \mathrm{Sp}_{2n}(\mathbb{Q})$, $k \in K_f$ and $m \in K_\infty$ such that $\tilde{x} = \gamma x k m$; so $x = \tilde{\gamma}\gamma x k\tilde{k}m\tilde{m}$. Then $k\tilde{k} = I$ since K_f is sufficiently small, so k defines a 1-cocycle.

Proposition 2.2 [Goresky and Tai 2003b]. *Conjugation by τ_0 on Sp_{2n} passes to an antiholomorphic involution $\eta : Y \to Y$ whose fixed point X is isomorphic to the finite disjoint union,*

$$X \cong \coprod_{\alpha \in H^1(\langle\tau_0\rangle, K_f)} X_\alpha$$

over cohomology classes α, where

$$X_\alpha = \mathrm{GL}_n(\mathbb{Q}) \backslash \mathrm{GL}_n(\mathbb{A}) / K_\alpha O(n)$$

is an arithmetic quotient of $\mathrm{GL}_n(\mathbb{R})$ and K_α is a certain[3] compact open subgroup of $\mathrm{GL}_n(\mathbb{A}_f)$. If $4 \mid N$ and if $K_f = \widehat{K}_N^0$ is the principal congruence subgroup of $\mathrm{Sp}_{2n}(\widehat{\mathbb{Z}})$ of level N then $K_\alpha = \widehat{K}_N'$ is independent of the cohomology class α, and X may be identified with the parameter space (or coarse moduli space) of principally polarized abelian varieties with real structure and level N structure. □

In this paper, by restricting to the case of ordinary abelian varieties, we make a first attempt at finding a finite field analog of Proposition 2.2.

2.3. The Siegel space $\mathfrak{h}_n$ admits another interesting antiholomorphic involution. In [Goresky and Tai 2003a] this involution is described on $\mathfrak{h}_2$ whose fixed point set is hyperbolic 3-space (cf. [Nygaard 1995]). After appropriate choice of level structure, it passes to an involution of the moduli space Y whose fixed point set is a union of arithmetic hyperbolic 3-manifolds which may be interpreted as constituting a coarse moduli space for abelian varieties with "antiholomorphic multiplication" by an order in an imaginary quadratic number field. A finite field analog for this result, along the same lines as the rest of this paper, which applies to the case of ordinary abelian varieties, is described in Section 12.

3. Deligne modules, polarizations and viable elements

3.1. ***Ordinary abelian varieties.*** Throughout this section we fix a finite field $k = \mathbb{F}_q$ of characteristic p. Let A/k be a dimension n abelian variety. Recall that A is *ordinary* if any of the following equivalent conditions is satisfied.

(1) If $\cdot p : A(\bar{k}) \to A(\bar{k})$ denotes the multiplication by p then its kernel has exactly p^n points.

(2) The local-local component of the p-divisible group $A(p^\infty) = \varprojlim A[p^r]$ is trivial.

[3]The class α vanishes in $H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\mathbb{A}_f))$ so there exists $h \in \mathrm{Sp}_{2n}(\mathbb{A}_f)$ such that $\alpha = [h^{-1}\tilde{h}]$. Then $K_\alpha = (h^{-1} K_f h) \cap \mathrm{Sp}_{2n}(\mathbb{A}_f)$ and right translation by h^{-1} maps $\mathrm{GL}_n(\mathbb{Q}) \backslash \mathrm{GL}_n(\mathbb{A}) / O(n).K_\alpha$ to Y.

(3) The middle coefficient of the characteristic polynomial h_A of the Frobenius endomorphism of A is not divisible by p.

(4) Exactly half of the roots of h_A in $\bar{\mathbb{Q}}_p$ are p-adic units.

3.2. Recall the basic definitions of [Deligne 1969]. A *Deligne module* of rank $2n$ over the field $k = \mathbb{F}_q$ of q elements is a pair (T, F) where T is a free $\mathbb{Z}$-module of dimension $2n$ and $F : T \to T$ is an endomorphism such that the following conditions are satisfied:

(1) The mapping F is semisimple and all of its eigenvalues in $\mathbb{C}$ have magnitude $\sqrt{q}$.

(2) Exactly half of the eigenvalues of F in $\bar{\mathbb{Q}}_p$ are p-adic units and half of the eigenvalues are divisible by q. (So $\pm\sqrt{q}$ is not an eigenvalue.)

(3) The middle coefficient of the characteristic polynomial of F is coprime to p.

(4) There exists an endomorphism $V : T \to T$ such that $FV = VF = q$.

A morphism $(T, F) \to (T', F')$ of Deligne modules is a group homomorphism $\phi : T \to T'$ such that $F'\phi = \phi F$.

3.3. Let $W(k)$ be the ring of (infinite) Witt vectors over k. Deligne [1969] chose an embedding

$$\varepsilon : W(\bar{k}) \to \mathbb{C} \tag{3.3.1}$$

("once and for all") which we henceforth refer to as *Deligne's embedding*. By a theorem of Serre and Tate [Drinfeld 1976; Katz 1981; Messing 1972; Nori and Srinivas 1987], the ordinary abelian variety A has a canonical lift $\bar{A}$ over $W(k)$ which, using (3.3.1), gives rise to a complex variety $A_{\mathbb{C}}$ over $\mathbb{C}$. Let $F \in \mathrm{Gal}(\bar{k}/k)$ denote the Frobenius. The geometric action of F on A lifts to an automorphism F_A on

$$T = T_A = H_1(A_{\mathbb{C}}, \mathbb{Z}).$$

Theorem 3.4 [Deligne 1969]. *The association $A \to (T_A, F_A)$, determined by Deligne's embedding* (3.3.1), *induces an equivalence of categories between the category of n-dimensional ordinary abelian varieties over $k = \mathbb{F}_q$ and the category of Deligne modules over k of rank $2n$.*

3.5. ***Endomorphism algebra and CM types.*** If A is an ordinary abelian variety over the finite field $k = \mathbb{F}_q$ of characteristic p, then A is $\mathbb{Q}$-isogenous $A \sim A_1 \times A_2 \times \cdots \times A_r$ to a product of ordinary abelian varieties over k such that:

(1) For $1 \le i \le r$, there exists a positive integer d_i and a simple ordinary abelian variety B_i over k and a $\mathbb{Q}$-isogeny $A_i \sim B_i^{d_i}$.

(2) $\mathrm{Hom}_{\mathbb{Q}}(B_i, B_j) = 0$ for $i \neq j$.

The endomorphism algebra $K_i = \mathrm{End}_{\mathbb{Q}}(B_i)$ is a CM field (that is, an imaginary quadratic extension of a maximal totally real subfield L_i) of degree $[K_i : \mathbb{Q}] = 2\dim(B_i)$. It is the center of the algebra $\mathrm{End}_{\mathbb{Q}}(A_i) \cong M_{d_i \times d_i}(K_i)$. The center of $\mathrm{End}_{\mathbb{Q}}(A)$ is therefore isomorphic to the CM algebra (that is, the product of CM fields) $K = K_1 \times \cdots \times K_r$.

A CM type Φ_i on K_i is a collection of embeddings $\phi : K_i \to \mathbb{C}$, one from each complex conjugate pair. It induces a real vector space isomorphism

$$\prod_{\phi \in \Phi_i} : K_i \otimes_{\mathbb{Q}} \mathbb{R} \to \mathbb{C}^{[K_i:\mathbb{Q}]/2}$$

which defines a complex structure on $K_i \otimes_{\mathbb{Q}} \mathbb{R}$. A CM type on K is a collection of nontrivial homomorphisms $\phi : K \to \mathbb{C}$, one from each complex conjugate pair, or equivalently, it is a choice of CM type for each K_i. Using Theorem 3.4 these statements become the following.

If (T, F) is the Deligne module corresponding to A then there is a decomposition

$$T_{\mathbb{Q}} = T \otimes_{\mathbb{Z}} \mathbb{Q} \cong T_{1,\mathbb{Q}} \oplus \cdots \oplus T_{r,\mathbb{Q}},$$

preserved by F, say $F = F_1 \oplus \cdots \oplus F_r$, and an isomorphism $\mathbb{Q}[F] \cong K_1 \times \cdots K_r$ of the center of $\mathrm{End}_{\mathbb{Q}}(T, F) = \mathrm{End}_F(T \otimes \mathbb{Q})$ with the CM algebra K. Then $T_{i,\mathbb{Q}}$ is a vector space of dimension d_i over the CM field $K_i = \mathbb{Q}[F_i]$. A CM type for $\mathbb{Q}[F]$ defines a complex structure on $T \otimes \mathbb{R}$. The minimum polynomial of (T, F) is the product of the minimum polynomials $h_i(x)$ of the $(T_{i,\mathbb{Q}}, F_i)$. It is an ordinary Weil q-polynomial (see Section A.1).

Deligne's embedding $\varepsilon : W(\bar{k}) \to \mathbb{C}$ induces a valuation val_p on $\overline{\mathbb{Q}} \subset \mathbb{C}$ extending the p-adic valuation of $W(k)$ (which explains the use of $W(\bar{k})$ rather than $W(k)$). This determines a canonical CM type for every Deligne module (and every CM algebra), which we refer to as *Deligne's CM type* as follows. If (T, F) is a Deligne module, define

$$\Phi_\varepsilon = \{\phi : \mathbb{Q}[F] \to \mathbb{C} \mid \mathrm{val}_p(\phi(F)) > 0\}. \tag{3.5.1}$$

Then Φ_ε is a CM type for the CM algebra $\mathbb{Q}[F]$. The resulting complex structure on $T \otimes_{\mathbb{Z}} \mathbb{R}$ is the unique complex structure such that the action of F is complex linear and such that $\mathrm{val}_p(\alpha) > 0$ for every eigenvalue α of F (see [Deligne 1969, p. 242]). It agrees with the complex structure on $T_0 A_{\mathbb{C}}$ in the case when $(T, F) = (T_A, F_A)$ is the Deligne module associated to an ordinary abelian variety A. The complex structure gives a Hodge structure on $T_0 A_{\mathbb{C}}$ which corresponds to an $\mathbb{R}$ homomorphism $\mathbb{S} = \mathrm{Res}_{\mathbb{C}/\mathbb{R}} \mathbb{G}_m \to \mathrm{GL}(T \otimes \mathbb{R})$.

3.6. Let (T, F) be a Deligne module. We are grateful to the referee for pointing out that not every CM type on $\mathbb{Q}[F]$ will arise as $\Phi_{\varepsilon'}$ for different embeddings $\varepsilon' : W(\bar{k}) \to \mathbb{C}$. Suppose as above that $K = \mathbb{Q}[F] \cong K_1 \times \cdots \times K_r$ is a decomposition

into a product of CM fields, with the corresponding decomposition $L = L_1 \times \cdots L_r$ of maximal totally real subfields. Each p-adic place of L_i splits in K_i but the prime p may ramify in L_i. A CM type arising from an embedding $\varepsilon' : W(\bar{k}) \to \mathbb{C}$ will correspond to a choice, for each i and for each p-adic place in L_i, of one of the two places in K_i over it. Let us say that such a CM type is *eligible*. Thus the total number of eligible CM types for $\mathbb{Q}[F]$ is 2^s where s is the number of p-adic places of $L_1 \times \cdots \times L_r$ (whereas the full number of CM types is 2^t where $t = \sum_i [L_i : \mathbb{Q}]$).

For Howe's theorem (Section 3.9) and Sections 6–10, we consider only eligible CM types on $\mathbb{Q}[F]$, and in fact, we use only Deligne's CM type Φ_ε. For Howe's theorem, this is crucial. However the results in Sections 6–9 are "linear algebra" statements that can be extended in a straightforward manner to include arbitrary CM types using Section 3.8 and parts (b) and (c) of Existence Lemma 3.10.

3.7. ***Polarizations.*** For a complex n-dimensional abelian variety X, a polarization may be considered to be a Hermitian form $H = R + i\omega$ defined on the (complex n-dimensional) tangent space T_0X, meaning that ω is a (real-valued) symplectic form on the underlying real vector space $(T_0X)_{\mathbb{R}}$ such that the inner product

$$R(x, y) = \omega(x, \sqrt{-1}.y)$$

is symmetric and positive definite. E. Howe [1995] defined the notion of a polarization of a Deligne module (T, F) in a similar way but the "positive definite" condition requires a replacement for the notion of multiplication by $\sqrt{-1}$.

If Φ is any CM type on $\mathbb{Q}[F]$, we will say (following [Howe 1995]) that an element $\iota \in \mathbb{Q}[F]$ is Φ-totally positive imaginary if $\phi(\iota)$ is a positive multiple of $\sqrt{-1}$ for all $\phi \in \Phi$. A polarization ω of the Deligne module (T, F) that is *positive with respect to the CM type* Φ is defined to be an alternating bilinear form $\omega : T \times T \to \mathbb{Z}$ such that

(0) $\omega(x, y) = -\omega(y, x)$ for all $x, y \in T$,

(1) $\omega : T_{\mathbb{Q}} \times T_{\mathbb{Q}} \to \mathbb{Q}$ is nondegenerate,

(2) $\omega(Fx, y) = \omega(x, Vy)$ for all $x, y \in T$,

and the following Φ-*positivity condition* (see Section 3.11) holds:

(∗) The bilinear form $R(x, y) = \omega(x, \iota y)$ is symmetric and positive definite for some (and hence any) totally Φ-positive imaginary element $\iota \in \mathbb{Q}[F]$.

Then we say that (T, F, ω) is *a* Φ-*positively polarized Deligne module*. (See also Section 3.11.)

Let us say that a symplectic form ω on a Deligne module (T, F) satisfying (1) and (2) above is a *polarization* if there exists a CM type Φ on $\mathbb{Q}[F]$ such that ω is a Φ-positive polarization on (T, F). (Most of this paper involves Deligne modules that are positively polarized with respect to Deligne's CM type Φ_ε.)

3.8. Suppose (T_1, F_1, ω_1) is Φ_1-positively polarized and that (T_2, F_2, ω_2) is Φ_2-positively polarized, where Φ_1, Φ_2 are CM types on $\mathbb{Q}[F_1]$, $\mathbb{Q}[F_2]$, respectively. If $g : (T_1, F_1) \to (T_2, F_2)$ is a morphism of Deligne modules that is compatible with the polarizations (meaning that $g^*(\omega_2) = \omega_1$) then it also follows that $g^*(\Phi_2) = \Phi_1$. Therefore we may speak of a morphism of polarized Deligne modules without necessarily referring to the CM type. (See also Existence Lemma 3.10).

Let S be a commutative ring with 1. An S-isogeny of Φ-positively polarized Deligne modules $\phi : (T_1, F_1, \omega_1) \to (T_2, F_2, \omega_2)$ is defined to be an S-isogeny (see Section 1.7) $\phi : (T_1, F_1) \to (T_2, F_2)$ for which there exists $c \in S^\times$ such that $\phi^*(\omega_2) = c\omega_1$, in which case c is called the *multiplier* of the isogeny ϕ. If $S \subset \mathbb{R}$ then any S-isogeny of Φ-positively polarized Deligne modules has positive multiplier.

3.9. ***Howe's theorem.*** If (T, F) is a Deligne module then the *dual* Deligne module $(\widehat{T}, \widehat{F})$ is defined by $\widehat{T} = \mathrm{Hom}(T, \mathbb{Z})$ and $\widehat{F}(\phi)(x) = \phi(Vx)$ for all $\phi \in \widehat{T}$.

Let A be an ordinary abelian variety with associated Deligne module (T_A, F_A) that is determined by the embedding ε of (3.3.1). Then there is a canonical isomorphism of Deligne modules, $(\widehat{T}_A, \widehat{F}_A) \cong (T_{\widehat{A}}, F_{\widehat{A}})$. Let $\omega : T_A \times T_A \to \mathbb{Z}$ be an alternating bilinear form that satisfies conditions (1) and (2) of Section 3.7. It induces an isomorphism

$$\lambda : (T_A \otimes \mathbb{Q}, F) \to (\widehat{T}_A \otimes \mathbb{Q}, \widehat{F})$$

and hence an isogeny $\lambda_A : A \to \widehat{A}$. Then [Howe 1995] proves that ω is positive *with respect to the CM type* Φ_ε if and only if λ_A is a polarization of the abelian variety A. Consequently, *the equivalence of categories in Theorem 3.4* (*which depends on the choice of* ε) *extends to an equivalence between the category of polarized* n*-dimensional abelian varieties over* $\mathbb{F}_q$ *with the category of* Φ_ε*-positively polarized Deligne modules* (*over* $\mathbb{F}_q$) *of rank* $2n$.

Existence Lemma 3.10 (see also Existence Lemma 4.4).

(**a**) *Let* $K = \mathbb{Q}[\pi]$ *be a CM field and let* Φ *be a CM type for* K. *Then there exists an integral symplectic form* $\omega : \mathcal{O}_K \times \mathcal{O}_K \to \mathbb{Z}$ *which satisfies the* Φ*-positivity condition* $(*)$ *of Section 3.7, that is, the bilinear form* $R(x, y) = \omega(x, \iota y)$ *is positive definite and symmetric, for any* Φ*-totally positive imaginary element* $\iota \in K$. *The form* ω *may be chosen so that* $\omega(ax, y) = \omega(x, \bar{a}y)$ *for any* $a \in K$ (*where bar denotes complex conjugation*); *hence* $\omega(\bar{x}, \bar{y}) = -\omega(x, y)$. *If* π *is an ordinary Weil* q*-number*[4] *then* $(\mathcal{O}_K, \pi, \omega)$ *is a* Φ*-positively polarized Deligne module.*

(**b**) *Let* (T, F) *be a Deligne module. For any CM type* Φ *on* $\mathbb{Q}[F]$, *there exists a* Φ*-positive polarization* ω *of* (T, F); *hence* (T, F, ω) *is a* Φ*-positively polarized Deligne module.*

[4] meaning that $\mathbb{Q}[\pi]$ has no real embeddings, that $\phi(\pi)\bar{\phi}(\pi) = q$ for each complex embedding $\phi : \mathbb{Q}[\pi] \to \mathbb{C}$, and that the middle coefficient of the characteristic polynomial of π is not divisible by p; see Appendix A.

(**c**) *Suppose (T, F) is a Deligne module and suppose $\omega : T \times T \to \mathbb{Z}$ is a symplectic form such that $\omega(Fx, y) = \omega(x, Vy)$. Then there exists a unique CM type Φ on $\mathbb{Q}[F]$ such that ω is Φ-positive; hence (T, F, ω) is a Φ-positively polarized Deligne module.*

Proof. A polarization that is positive with respect to Φ and compatible with complex conjugation is described in [Shimura and Taniyama 1961, §6.2] and [Shimura 1998, §6.2]; see also [Milne 2005, Proposition 10.2, p. 335]; we repeat the definition here. Let $\alpha \in \mathcal{O}_K$ be totally Φ-positive imaginary and for all $x, y \in K$ set

$$\omega_K(x, y) = \mathrm{Trace}_{K/\mathbb{Q}}(\alpha x.\bar{y}).$$

Then $\omega_K : \mathcal{O}_K \times \mathcal{O}_K \to \mathbb{Z}$ is antisymmetric, the bilinear form $R(x, y) = \omega_K(x, \alpha y)$ is symmetric and positive definite, $\omega_K(\pi x, \pi y) = q\omega_K(x, y)$, and $\omega_K(\bar{x}, \bar{y}) = -\omega_K(x, y)$.

Part (b) follows from part (a) by decomposition into simple Deligne modules.

For part (c) we may also suppose (T, F) is $\mathbb{Q}$-simple and $\omega : T \times T \to \mathbb{Z}$ is alternating and nondegenerate over $\mathbb{Q}$ with $\omega(Fx, y) = \omega(x, Vy)$. A choice of an F-cyclic vector gives an isomorphism of $\mathbb{Q}[F]$-modules, $T \otimes \mathbb{Q} \cong \mathbb{Q}[F]$. Using this isomorphism, the mapping $x \mapsto \omega(x, 1)$ is $\mathbb{Q}$-linear, so is given by $\mathrm{Trace}_{K/\mathbb{Q}}(\alpha x)$ for some uniquely determined $\alpha \in K$. It follows that $\bar{\alpha} = -\alpha$ and

$$\omega(x, y) = \mathrm{Trace}_{K/\mathbb{Q}}(\alpha x \bar{y}).$$

This element α determines a CM type $\Phi = \Phi_\alpha$ for K: for any embedding $K \to \mathbb{C}$ the image of α is purely imaginary so there is a unique choice ϕ from each pair of complex conjugate embeddings such that $\phi(\alpha)$ is positive imaginary. It is easy to check that ω is Φ_α-positive. For uniqueness, if $\beta \in \mathbb{Q}[F]$ is any other element such that $(x, y) \to \omega(x, \beta y)$ is symmetric and positive definite then $\bar{\beta} = -\beta$ so $\phi(\beta)$ is purely imaginary for every $\phi \in \Phi_\alpha$. Moreover,

$$\omega(x, \beta x) = \sum_{\phi \in \Phi_\alpha} \phi(\alpha\bar{\beta}x\bar{x}) + \overline{\phi}(\alpha\bar{\beta}x\bar{x}) = 2\sum_{\phi \in \Phi_\alpha} \phi(\alpha\bar{\beta}x\bar{x}) > 0$$

for all $x \in \mathbb{Q}[F]^\times$. This implies that $\phi(\alpha)\phi(\bar{\beta}) = -\phi(\alpha)\phi(\beta) > 0$ for each $\phi \in \Phi_\alpha$; hence $\phi(\beta)$ is also positive imaginary, that is, $\Phi_\alpha = \Phi_\beta$. □

3.11. ***Viable elements.*** Let $\gamma_0 \in \mathrm{GSp}_{2n}(\mathbb{Q})$ be a semisimple element whose characteristic polynomial is an ordinary Weil q-polynomial. If $\gamma, \gamma_0 \in \mathrm{GSp}_{2n}(\mathbb{Q})$ are stably conjugate[5] then conjugation defines a unique isomorphism $\mathbb{Q}[\gamma] \cong \mathbb{Q}[\gamma_0]$. Let $\mathcal{C} \subset \mathrm{GSp}_{2n}(\mathbb{Q})$ be the stable conjugacy class of γ_0 and let Φ be a CM type on the CM algebra $K = K(\mathcal{C}) = \mathbb{Q}[\gamma_0]$. An element $\gamma \in \mathcal{C}$ will be said to be Φ-*viable* if

[5]Meaning that there exists $g \in \mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$ such that $\gamma = g^{-1}\gamma_0 g$.

the pair (γ, ω_0) satisfies the following positivity condition (where ω_0 is the standard symplectic form on $\mathbb{Q}^{2n}$):

(∗∗) The bilinear form $R(x, y) = \omega_0(x, \iota y)$ is symmetric and positive definite on $\mathbb{Q}^{2n}$ for any totally Φ-positive imaginary element $\iota \in \mathbb{Q}[\gamma]$.

Proposition 3.12. (see also Existence Lemma 4.4.) *Let $\mathcal{C} \subset \mathrm{GSp}_{2n}(\mathbb{Q})$ be a stable conjugacy class of semisimple elements whose characteristic polynomial is an ordinary Weil q polynomial and let Φ be a CM type on the associated CM algebra $K = K(\mathcal{C})$.*

(1) *An element $\gamma \in \mathcal{C}$ is Φ-viable if and only if there exists a Φ-positively polarized Deligne module of the form (L, γ, ω_0) for some lattice $L \subset \mathbb{Q}^{2n}$.*

(2) *The set of Φ-viable elements in $\mathcal{C}$ is nonempty and forms a unique $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugacy class*[6] *within $\mathcal{C}$.*

Proof. For part (1), given a Φ-viable element $\gamma \in \mathcal{C}$, we may reduce to the case that the characteristic polynomial of γ is irreducible and $K = \mathbb{Q}[\gamma]$ is a CM field. We need to construct a lattice $L \subset \mathbb{Q}^{2n}$ which is preserved by γ and by $q\gamma^{-1}$, such that the symplectic form ω_0 take integer values on L.

Let $v_0 \in \mathbb{Q}^{2n}$ be a cyclic vector for the action of γ, that is, a generator of $\mathbb{Q}^{2n}$ as a one-dimensional $K = \mathbb{Q}[\gamma]$ module and let $\psi : K \to \mathbb{Q}^{2n}$ be the unique K-equivariant mapping such that $\psi(1) = v_0$.

Let $\omega_K = \psi^*(\omega_0)$. Then $\omega_K(x, 1)$ is linear in x so it is given by $\mathrm{Trace}_{K/\mathbb{Q}}(\alpha x)$ for some unique element $\alpha \in K$. It follows that $\omega_K(x, y) = \mathrm{Trace}_{K/\mathbb{Q}}(\alpha x \bar{y})$, that $\bar{\alpha} = -\alpha$ and that α is totally Φ-positive imaginary. If we change v_0 to $x.v_0$ (with $x \in \mathbb{Q}[\gamma]$) then α changes to $x\bar{x}\alpha$. It follows that we may choose v_0 so that $\alpha \in K$ is an algebraic integer. Therefore multiplication by α preserves $\mathcal{O}_K$ so ω_K is integer valued on $\mathcal{O}_K$. Hence, we may take $L = \psi(\mathcal{O}_K)$.

For part (2), by Existence Lemma 3.10(c), there exists a Φ-viable element $\gamma_0 \in \mathcal{C}$, and there exists a Φ-positively polarized Deligne module of the form $(L_0, \gamma_0, \omega_0)$ (where $L_0 \subset \mathbb{Q}^{2n}$). Now let $\gamma \in \mathcal{C}$. We must show that γ is Φ-viable if and only if it is $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugate to γ_0.

First, suppose that γ is Φ-viable, and hence (L, γ, ω_0) is a Φ-positively polarized Deligne module, for some lattice $L \subset \mathbb{Q}^{2n}$. Since γ, γ_0 have the same characteristic polynomial, there exists $\phi \in \mathrm{GL}_{2n}(\mathbb{Q})$ with $\gamma_0 = \phi^{-1}\gamma\phi$ which therefore induces an identification $\mathbb{Q}[\gamma] \cong \mathbb{Q}[\gamma_0]$. So $(\phi^{-1}(L), \phi^*(\gamma) = \gamma_0, \phi^*(\omega_0))$ is a Φ-positively polarized Deligne module. Choose $c \in \mathbb{Q},\ c > 0$, such that $c\phi^*(\omega_0)$ takes integer values on L_0. Then $c\phi^*(\omega_0)$ is a second polarization of the Deligne module (L_0, γ_0). By Lemma C.1 there is an $\mathbb{R}$-isogeny $\psi : (L_0, \gamma_0, \omega_0) \to (L_0, \gamma_0, c\phi^*(\omega_0))$ with multiplier equal to 1, which implies that $\psi^*\phi^*(c\omega_0) = \omega_0$. Thus, conjugation

[6] Meaning the intersection of an $\mathrm{Sp}_{2n}(\mathbb{R})$ conjugacy class with $\mathcal{C}$.

by $\phi \circ \psi$ takes γ_0 to γ, and $\phi \circ \psi \in \mathrm{GSp}_{2n}(\mathbb{R})$ has multiplier $c > 0$. Therefore conjugation by $1/\sqrt{c}\,\phi \circ \psi \in \mathrm{Sp}_{2n}(\mathbb{R})$ also takes γ_0 to γ. The converse is similar (but easier). □

4. Real structures

Definition 4.1. Fix a Deligne module (T, F) over $k = \mathbb{F}_q$ of dimension $2n$. A *real structure* on (T, F) is a $\mathbb{Z}$-linear homomorphism $\tau : T \to T$ such that $\tau^2 = I$ and such that $\tau F \tau^{-1} = V$. A (real) morphism $\phi : (T, F, \tau) \to (T', F', \tau')$ of Deligne modules with real structures is a group homomorphism $\phi : T \to T'$ such that $\phi F = F'\phi$ and $\phi\tau = \tau'\phi$. A real structure τ is compatible with a polarization $\omega : T \times T \to \mathbb{Z}$ if, for all $x, y \in T$,

$$\omega(\tau x, \tau y) = -\omega(x, y). \tag{4.1.1}$$

Let $N \geq 1$ and assume $p \nmid N$. A (*principal*) *level N structure* on (T, F) is an isomorphism $\beta : T/NT \to (\mathbb{Z}/N\mathbb{Z})^{2n}$ such that $\beta \circ \bar{F} = \beta$ where $\bar{F} = F \pmod N$. (If a level N structure exists, it implies that $F \equiv I \pmod N$, which places further restrictions on N.) A level N structure is said to be compatible with a polarization $\omega : T \times T \to \mathbb{Z}$ if $\beta_*(\omega) = \bar{\omega}_0$ is the reduction modulo N of the standard symplectic form (see Section 1.7).

If (T, F, τ) is a Deligne module with real structure then a level N structure β on (T, F) is *compatible with* τ if $\beta_*(\tau) = \bar{\tau}_0$ is the reduction modulo N of the standard involution (see Section 1.7 and Section D.3). A necessary condition for the existence of a level N structure that is compatible with τ is that $p \equiv 1 \pmod N$, which also implies that $V \equiv I \pmod N$; see Section 5.1.

4.2. In Theorem 7.1 we will prove (for q, N coprime) that there are finitely many isomorphism classes of principally (Φ_ε-positively) polarized Deligne modules of rank $2n$ over $\mathbb{F}_q$ with real structure and with principal level N structure. In Section 11.1 we add a few remarks concerning the fixed point lattice T^τ (or "real sublattice") of a Deligne module (T, F) with real structure τ.

Lemma 4.3. *The category of Deligne modules* (*resp. polarized Deligne modules*) *with real structure, up to* $\mathbb{Q}$*-isogeny is semisimple. If* (T, F, τ) *is* $\mathbb{Q}$*-simple then either* (*a*) $T_{\mathbb{Q}} = T \otimes \mathbb{Q}$ *is a simple* $\mathbb{Q}[F]$ *module or* (*b*) *there exists a simple* $\mathbb{Q}[F]$ *module* W *so that* $T_{\mathbb{Q}} \cong W \oplus \tau(W)$.

Proof. The proof is more or less standard. For the first statement, it suffices to check complete reducibility. Let (T, F, τ) be a Deligne module, and let (T_1, F, τ) be a submodule. Since F is semisimple the ring $\mathbb{Q}[F]$ is isomorphic to a product of number fields. It follows that (T, F, τ) decomposes into a sum of modules over these constituent fields. So we may assume that $\mathbb{Q}[F]$ is a field. Set $W = T \otimes \mathbb{Q}$ and

let $W_1 = T_1 \otimes \mathbb{Q}$. Choose any decomposition of W into simple $\mathbb{Q}[F]$-submodules so that W_1 is a summand. The resulting projection $\pi : W \to W_1$ is $\mathbb{Q}[F]$-equivariant. Let $e = \pi + \tau\pi\tau : W \to W_1$. Then e is surjective (since its restriction to W_1 coincides with multiplication by 2) and $W_1' := \ker(e)$ is preserved by F and by τ. Thus, the decomposition $W = W_1 \oplus W_1'$ is preserved by F and by τ. For any choice of lattice $T_1' \subset W_1'$ preserved by F and τ the module $(T_1 \oplus T_1', F, \tau)$ is $\mathbb{Q}$-isogenous to (T, F, τ). The statement about simple modules follows.

Similarly suppose (T, F, ω, τ) is a Deligne module with real structure and Φ-positive polarization with respect to a choice Φ of CM type on $\mathbb{Q}[F]$. Let $W = T \otimes \mathbb{Q}$ and suppose that $W_1 \subset W$ is a subspace preserved by F and by τ. Set $F_1 = F|W_1$. It follows that

(1) the restriction of ω to W_1 is nondegenerate and is Φ_1-positive, where Φ_1 is the CM type on $\mathbb{Q}[F_1]$ that is induced from Φ,

(2) the subspace $W_2 = \{y \in W \mid \omega(w, y) = 0 \text{ for all } w \in W_1\}$ is also preserved by F and by τ and it is Φ_2-positively polarized by the restriction $\omega \mid W_2$ where Φ_2 is the CM type induced from Φ on $\mathbb{Q}[F_2]$ (where $F_2 = F|W_2$), and

(3) the module W decomposes as an orthogonal sum $W = W_1 \oplus W_2$. □

Existence Lemma 4.4. *The Φ-positively polarized Deligne module $(\mathcal{O}_K, \pi, \omega)$ defined in part (a) of Existence Lemma 3.10 admits a canonical real structure given by complex conjugation. Statement (b) of Existence Lemma 3.10 remains true if the Deligne module (T, F) is replaced by a Deligne module with real structure (T, F, τ), in which case the resulting polarization ω is compatible with the real structure. Statement (c) remains true if the Deligne module (T, F) has a real structure.*

Let $\mathcal{C} \subset \mathrm{GSp}_{2n}(\mathbb{Q})$ be a stable conjugacy class as in Proposition 3.12, and let Φ be a CM type on the CM algebra $K = K(\mathcal{C})$. Let $\gamma \in \mathcal{C}$ be Φ-viable and also q-inversive (see Section 5). Then there exists a lattice $L \subset \mathbb{Q}^{2n}$ that is preserved by γ, $q\gamma^{-1}$, and by the standard involution τ_0 so that $(L, \gamma, \omega_0, \tau_0)$ is a Φ-positively polarized Deligne module with real structure.

Proof. The first three statements are easy to verify. The last statement follows from the same proof as that of Existence Lemma 3.10 and Proposition 3.12, using Lemma 4.3 to reduce to the simple case. □

5. q-inversive elements

5.1. Let R be an integral domain. Let us say that an element $\gamma \in \mathrm{GSp}_{2n}(R)$ is *q-inversive* if it is semisimple, has multiplier q and if[7]

$$\tau_0 \gamma \tau_0^{-1} = q\gamma^{-1},$$

[7] Compare the equation $\tau F \tau^{-1} = V$ of Definition 4.1.

or equivalently if $\gamma = \left(\begin{smallmatrix} A & B \\ C & {}^tA \end{smallmatrix}\right) \in \mathrm{GSp}_{2n}(R)$ and B, C are symmetric, and $A^2 - BC = qI$. It follows that $B\,{}^tA = AB$ and $CA = {}^tAC$. In Lemma 6.3 below, it is explained that the endomorphism F of a polarized Deligne module with real structure may be represented by a q-inversive element.

Lemma 5.2. *Let $\gamma = \left(\begin{smallmatrix} A & B \\ C & {}^tA \end{smallmatrix}\right) \in \mathrm{GSp}_{2n}(\mathbb{Q})$ be q-inversive. Then the following statements are equivalent.*

(1) *The matrices A, B, and C are nonsingular.*

(2) *The element γ has no eigenvalues in the set $\{\pm\sqrt{q}, \pm\sqrt{-q}\}$.*

If these properties hold then the matrix A is semisimple, and the characteristic polynomial of A is $h(2x)$, where $h(x)$ is the real counterpart (see Section A.2) *to $g(x)$, the characteristic polynomial of γ. If $g(x)$ is also an ordinary Weil q-polynomial then $p \nmid \det(A)$ and every eigenvalue β of A satisfies*

$$|\beta| < \sqrt{q}. \tag{5.2.1}$$

Conversely, let $A \in \mathrm{GL}_n(\mathbb{Q})$ be semisimple and suppose that its eigenvalues $\beta_1, \ldots, \beta_n$ (not necessarily distinct) are totally real and that $|\beta_r| < \sqrt{q}$ for $1 \le r \le n$. Then for any symmetric nonsingular matrix $C \in \mathrm{GL}_n(\mathbb{Q})$ such that ${}^tAC = CA$, the element

$$\gamma = \begin{pmatrix} A & (A^2 - qI)C^{-1} \\ C & {}^tA \end{pmatrix} \in \mathrm{GSp}_{2n}(\mathbb{Q}) \tag{5.2.2}$$

is q-inversive and its eigenvalues are the Weil q-numbers:

$$\alpha_r = \beta_r \pm \sqrt{\beta_r^2 - q}, \qquad 1 \le r \le n. \tag{5.2.3}$$

Proof. These statements are direct consequences of the following observation: if $w = \left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)$ is an eigenvector of γ with eigenvalue λ then

(a) $\tau_0(w) = \left(\begin{smallmatrix} -u \\ v \end{smallmatrix}\right)$ is an eigenvector of γ with eigenvalue q/λ,

(b) u is an eigenvector of A with eigenvalue $\frac{1}{2}\left(\lambda + \frac{q}{\lambda}\right)$,

(c) v is an eigenvector of ${}^tA$ with eigenvalue $\frac{1}{2}\left(\lambda + \frac{q}{\lambda}\right)$. □

5.3. ***Joint signature.*** Let $E_n(\mathbb{R})$ denote the set of pairs (A, C) where $A \in \mathrm{GL}_n(\mathbb{R})$ is semisimple with all eigenvalues real, where $C \in \mathrm{GL}_n(\mathbb{R})$ is symmetric, and where A is self adjoint with respect to the inner product $\langle \cdot, \cdot \rangle_C$ defined by C, that is, ${}^tAC = CA$. If $\beta \neq \mu$ are eigenvalues of A then the eigenspaces V_β and V_μ are orthogonal with respect to $\langle \cdot, \cdot \rangle_C$. Therefore $\langle \cdot, \cdot \rangle_C$ decomposes as a direct sum of bilinear forms $\bigoplus_{\beta \in \mathrm{Spec}(A)} \langle \cdot, \cdot \rangle_\beta$ with respect to the eigenspace decomposition $\mathbb{R}^n = \bigoplus_{\beta \in \mathrm{Spec}(A)} V_\beta$ where $\mathrm{Spec}(A) \subset \mathbb{R}$ denotes the spectrum of A. Define

$$\mathrm{sig}(A; C) = \{\mathrm{sig}\ \langle \cdot, \cdot \rangle_\beta\}_{\beta \in \mathrm{Spec}(A)}$$

to be the ordered collection of signatures of each of these bilinear forms. The elements of $\mathrm{sig}(A; C)$ correspond to the class in the Galois cohomology set

$$H^1(\mathbb{C}/\mathbb{R}, Z_A \cap O(C))$$

of the centralizer of A intersected with the orthogonal group of C.

The group $\mathrm{GL}_n(\mathbb{R})$ acts on $E_n(\mathbb{R})$ by

$$X.(A, C) = (XAX^{-1}, {}^tX^{-1}CX^{-1}).$$

Two elements (A, C) and (A', C') are in the same orbit if and only if A, A' have the same characteristic polynomial and $\mathrm{sig}(A, C) = \mathrm{sig}(A', C')$. In fact, the stabilizer of A in $\mathrm{GL}_n(\mathbb{R})$ is $\prod_{\beta\in\mathrm{Spec}(A)} \mathrm{GL}(V_\beta)$ and within each V_β the congruence class[8] of C_β is determined by the signature of $\langle \cdot, \cdot \rangle_\beta$.

5.4. ***Conjugacy of q-inversive elements.*** In this section we consider GL_n versus Sp_{2n} conjugacy of q-inversive elements. Let $L \supset \mathbb{Q}$ be a field. The subgroup of $\mathrm{GSp}_{2n}(L)$ that is fixed under conjugation by the standard involution τ_0 is denoted $\mathrm{GL}_n^*(L)$, and it is the image of the *standard embedding*

$$\delta : L^\times \times \mathrm{GL}_n(L) \to \mathrm{GSp}_{2n}(L); \quad \delta(\lambda, x) = \begin{pmatrix} \lambda X & 0 \\ 0 & {}^tX^{-1} \end{pmatrix}.$$

(For $\lambda = 1$ we use the same notation $\delta : \mathrm{GL}_n \to \mathrm{Sp}_{2n}$.) Say that two elements of GSp_{2n} are GL_n^* (resp. GL_n)-conjugate if the conjugating element lies in the image of δ (resp. $\delta(1 \times \mathrm{GL}_n)$). Then GL_n^*-conjugation preserves q-inversive elements.

Proposition 5.5. *Let* $\gamma_1, \gamma_2 \in \mathrm{GSp}_{2n}(\mathbb{Q})$ *be q-inversive, say* $\gamma_i = \left(\begin{smallmatrix} A_i & B_i \\ C_i & {}^tA_i \end{smallmatrix}\right)$. *Then*

$$\begin{aligned} \gamma_1, \gamma_2 \textit{ are } \mathrm{GSp}_{2n}(\bar{\mathbb{Q}})\textit{-conjugate} &\iff A_1, A_2 \textit{ are } \mathrm{GL}_n(\mathbb{Q})\textit{-conjugate} \\ &\iff \gamma_1, \gamma_2 \textit{ are } \mathrm{GL}_n(\bar{\mathbb{Q}})\textit{-conjugate} \\ \gamma_1, \gamma_2 \textit{ are } \mathrm{Sp}_{2n}(\mathbb{R})\textit{-conjugate} &\iff \gamma_1, \gamma_2 \textit{ are } \mathrm{GL}_n(\mathbb{R})\textit{-conjugate} \\ &\iff A_1, A_2 \textit{ are } \mathrm{GL}_n(\mathbb{Q})\textit{-conjugate and} \\ &\qquad \mathrm{sig}(A_1; C_1) = \mathrm{sig}(A_2; C_2). \end{aligned}$$

Proof. Conjugacy by $\mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$ is the same as conjugacy by $\mathrm{Sp}_{2n}(\bar{\mathbb{Q}})$ and, among semisimple elements, is determined by the characteristic polynomial. Lemma 5.2 gives that the characteristic polynomial of γ_i determines that of A_i and vice versa. Conjugacy of semisimple rational matrices A_1, A_2 is determined by the characteristic polynomial. This proves the first statement. Using $\mathrm{GL}_n(\bar{\mathbb{Q}})$ it is possible to diagonalize A_i and to reduce C_i to the identity, which proves the second statement.

[8] Symmetric matrices S and T are *congruent* if there exists a matrix X so that $T = XS\,{}^tX$.

For the third implication ($\Leftarrow$), equality of the signatures guarantees the existence of $X \in \mathrm{GL}_n(\mathbb{R})$ so that $X.(A_1, C_1) = (A_2, C_2) \in E_n(\mathbb{R})$ as explained in Section 5.3. Then γ_1, γ_2 are conjugate by $\left(\begin{smallmatrix} X & 0 \\ 0 & {}^tX^{-1} \end{smallmatrix}\right) \in \mathrm{Sp}_{2n}(\mathbb{R})$.

Now suppose that γ_1, γ_2 are $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugate. Then A_1, A_2 are $\mathrm{GL}_n(\mathbb{Q})$-conjugate since they have the same characteristic polynomial, so we need to show that $\mathrm{sig}(A_1; C_1) = \mathrm{sig}(A_2; C_2)$ or, equivalently, that γ_1, γ_2 are conjugate by an element of $\delta(\mathrm{GL}_n(\mathbb{R}))$. As in Section 5.3, conjugating by elements of $\delta(\mathrm{GL}_n(\mathbb{R}))$ and by decomposing with respect to the eigenspace decompositions of A_1, A_2, we may reduce to the case that $A_1 = A_2 = \lambda I_n$, and that C_1, C_2 are diagonal matrices consisting of ± 1.

So, let us assume that $C_1 = I_r$ consists of r copies of $+1$ and $n-r$ copies of -1 along the diagonal, and that $C_2 = I_s$. This determines $B_1 = dI_r$ and $B_2 = dI_s$ where $d = \lambda^2 - q$. Assuming that γ_1, γ_2 are $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugate, we need to prove that $r = s$.

Suppose $h = \left(\begin{smallmatrix} X & Y \\ Z & W \end{smallmatrix}\right) \in \mathrm{Sp}_{2n}(\mathbb{R})$ and $\gamma_2 = h\gamma_1 h^{-1}$. Subtracting $\lambda I_{2n\times 2n}$ from both sides of this equation leaves

$$\begin{pmatrix} X & Y \\ Z & W \end{pmatrix}\begin{pmatrix} 0 & dI_r \\ I_r & 0 \end{pmatrix} = \begin{pmatrix} 0 & dI_s \\ I_s & 0 \end{pmatrix}\begin{pmatrix} X & Y \\ Z & W \end{pmatrix} \tag{5.5.1}$$

or $W = I_s X I_r$ and $Z = d^{-1} I_s Y I_r$. Let $H = X + 1/\sqrt{d}\, Y I_r \in \mathrm{GL}_{2n}(\mathbb{C})$. Then

$$H I_r\, {}^t\bar{H} = \left(X + \frac{1}{\sqrt{d}} Y I_r\right) I_r\, {}^t\left(X - \frac{1}{\sqrt{d}} Y I_r\right) = I_s$$

for the real part of this equation comes from $X\,{}^tW - Y\,{}^tZ = I$ (see (B.1.3)) and the imaginary part follows similarly because $h \in \mathrm{Sp}_{2n}(\mathbb{R})$. But I_r and I_s are Hermitian matrices so this equation implies that their signatures are equal, that is, $r = s$. □

5.6. Let $h(x) \in \mathbb{Z}[x]$ be a real, ordinary Weil q-polynomial (see Appendix A); that is:

(h1) $h(0)$ is relatively prime to q.

(h2) The roots $\beta_1, \beta_2, \ldots, \beta_n$ of h are totally real and $|\beta_i| < 2\sqrt{q}$ for $1 \le i \le n$.

Let $\mathcal{S}(h)$ be the algebraic variety, defined over $\mathbb{Q}$, consisting of all pairs (A_0, C) where $A_0, C \in \mathrm{GL}_n$, where A_0 is semisimple and its characteristic polynomial is equal to $h(2x)$, where C is symmetric and ${}^tA_0 C = C A_0$. As in Lemma 5.2, there is a natural mapping

$$\theta : \mathcal{S}(h) \to \mathrm{GSp}_{2n}, \quad (A_0, C) \mapsto \begin{pmatrix} A_0 & B \\ C & {}^tA_0 \end{pmatrix}, \tag{5.6.1}$$

where $B = (A_0^2 - qI)C^{-1}$. The image $\theta(\mathcal{S}(h)_{\mathbb{Q}})$ of the set of rational elements consists of all q-inversive elements whose characteristic polynomial is the ordinary

Weil q-polynomial $p(x) = x^n h(x + q/x)$ (see Appendix A). The image of θ is preserved by the action of GL_n, which corresponds to the action

$$X.(A_0, C) = (XA_0X^{-1}, {}^tX^{-1}CX^{-1}) \tag{5.6.2}$$

for $X \in \mathrm{GL}_n$. In the notation of Lemma 5.2, the orbits of $\mathrm{GL}_n(\mathbb{R})$ on $\mathcal{S}(h)_{\mathbb{R}}$ are uniquely indexed by the values $\mathrm{sig}(A_0; C) = \{\mathrm{sig}(C_\beta)\}$ of the signature of each of the quadratic forms C_β on the eigenspace V_β, as β varies over the distinct roots of $h(x)$. By abuse of terminology we shall refer to the rational elements in the $\mathrm{GL}_n(\mathbb{R})$ orbit of $(A_0, C) \in \mathcal{S}(h)_{\mathbb{Q}}$ as the "$\mathrm{GL}_n(\mathbb{R})$-orbit containing (A_0, C)".

5.7. Let $(A_0, C_0) \in \mathcal{S}(h)_{\mathbb{Q}}$ and set $\gamma = \theta(A_0) \in \mathrm{GSp}_{2n}(\mathbb{Q})$ as in (5.6.1). The algebra $K = \mathbb{Q}[\gamma]$ is isomorphic to a product of CM fields (see Section 3.5). Fix a CM type Φ for K.

Recall from Proposition 3.12 (resp. Existence Lemma 4.4) that in order for the pair (γ, ω_0) (resp. the triple $(\gamma, \omega_0, \tau_0)$) to give rise to a Φ-polarized Deligne module (resp. Φ-polarized Deligne module with real structure), it is necessary and sufficient that γ should be Φ-viable.

Proposition 5.8. *Fix $h(x)$ and Φ as in Sections 5.6 and 5.7. For any semisimple matrix $A_0 \in \mathrm{GL}_n(\mathbb{Q})$ with characteristic polynomial equal to $h(2x)$ there exists a symmetric nonsingular element $C_0 \in \mathrm{GL}_n(\mathbb{Q})$ so that $(A_0, C_0) \in \mathcal{S}(h)_{\mathbb{Q}}$ and so that $\gamma_0 = \theta(A_0, C_0) \in \mathrm{GSp}_{2n}(\mathbb{Q})$ is Φ-viable. For every $(A, C) \in \mathcal{S}(h)_{\mathbb{Q}}$ the corresponding element $\gamma = \theta(A, C)$ is Φ-viable if and only if it is $\delta(\mathrm{GL}_n(\mathbb{R}))$-conjugate to γ_0.*

Proof. Given A_0 we need to prove the existence of $C_0 \in \mathrm{GL}_n(\mathbb{Q})$ such that $(A_0, C_0) \in \mathcal{S}(h)$ is Φ-viable. By Existence Lemma 4.4 there is a Φ-polarized Deligne module with real structure (T, F, ω, τ), whose characteristic polynomial is $p(x)$. Use Proposition B.4 to choose a basis $h : T \otimes \mathbb{Q} \xrightarrow{\sim} \mathbb{Q}^{2n}$ so that $h(T) \subset \mathbb{Q}^{2n}$ is a lattice, so that $h_*(\omega) = \omega_0$ and that $h_*(\tau) = \tau_0$ in which case the mapping F becomes a matrix $\gamma = \left(\begin{smallmatrix} A & B \\ C & {}^tA \end{smallmatrix}\right)$. It follows that γ is viable and that the characteristic polynomial of A is equal to that of A_0. So there exists $X \in \mathrm{GL}_n(\mathbb{Q})$ satisfying $A_0 = XAX^{-1}$. Define $C_0 = {}^tX^{-1}CX^{-1}$ so that $(A_0, C_0) = X \cdot (A, C)$. Then $\gamma_0 = \theta(A_0, B_0) = \delta(X)\gamma\delta(X)^{-1}$ is q-inversive, its characteristic polynomial is $p(x)$, and by Proposition 3.12, it is viable.

For the second statement, γ is Φ-viable if and only if it is $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugate to γ_0, by Proposition 3.12. This holds if and only if it is $\delta(\mathrm{GL}_n(\mathbb{R}))$-conjugate to γ_0, by Proposition 5.5. □

5.9. ***Remark.*** In the notation of the preceding paragraph, $\gamma = \theta(A, C)$ is Φ-viable if and only if $\mathrm{sig}(A, C) = \mathrm{sig}(A_0, C_0)$. If the roots of $h(x)$ are distinct then the CM field $\mathbb{Q}[\gamma]$ has 2^n different CM types, corresponding to the 2^n possible values of

$\mathrm{sig}(A, C)$ (that is, an ordered n-tuple of ± 1). However, if $h(x)$ has repeated roots then there exist elements $(A, C) \in \mathcal{S}(h)_{\mathbb{Q}}$ such that $\gamma = \theta(A, C)$ is not viable for any choice Φ of CM type on $\mathbb{Q}[\gamma]$.

6. $\bar{\mathbb{Q}}$-isogeny classes

The first step in counting the number of (principally polarized) Deligne modules (with or without real structure) is to identify the set of $\bar{\mathbb{Q}}$ isogeny classes of such modules, following the method of Kottwitz [1990]. Throughout this and subsequent sections we shall only consider polarizations that are positive with respect to the CM type Φ_ε as described in Section 3.7.

Lemma 6.1. *For $i = 1, 2$, let (T_i, F_i) be a Deligne module with (Φ_ε-positive) polarization ω_i. Let p_i be the characteristic polynomial of F_i. Then the following statements are equivalent.*

(1) *The characteristic polynomials are equal*: $p_1(x) = p_2(x)$.

(2) *The Deligne modules (T_1, F_1) and (T_2, F_2) are $\mathbb{Q}$-isogenous.*

(3) *The Deligne modules (T_1, F_1) and (T_2, F_2) are $\bar{\mathbb{Q}}$-isogenous.*

(4) *The polarized Deligne modules (T_1, F_1, ω_1) and (T_2, F_2, ω_2) are $\bar{\mathbb{Q}}$-isogenous.*

For $i = 1, 2$, suppose the polarized Deligne module (T_i, F_i, ω_i) admits a real structure τ_i. Then (1), (2), (3), (4) *are also equivalent to the following statements*:

(5) *The real Deligne modules (T_1, F_1, τ_1) and (T_2, F_2, τ_2) are $\mathbb{Q}$-isogenous.*

(6) *The real Deligne modules (T_1, F_1, τ_1) and (T_2, F_2, τ_2) are $\bar{\mathbb{Q}}$-isogenous.*

(7) *The real polarized Deligne modules $(T_1, F_1, \omega_1, \tau_1)$ and $(T_2, F_2, \omega_2, \tau_2)$ are $\bar{\mathbb{Q}}$-isogenous.*

Proof. Clearly, (4)⇒(3)⇒(1) and (2)⇒(1). The implication (1)⇒(2) is a special case of a theorem of Tate, but in our case it follows immediately from the existence of rational canonical form (see, for example, [Knapp 2006, p. 443]) that is, by decomposing $T_i \otimes \mathbb{Q}$ into F_i-cyclic subspaces ($i = 1, 2$) and mapping cyclic generators in T_1 to corresponding cyclic generators in T_2.

The proof that (2)⇒(4) is a special case from [Kottwitz 1990, p. 206], which proceeds as follows. Given $\phi : (T_1 \otimes \mathbb{Q}, F_1) \to (T_2 \otimes \mathbb{Q}, F_2)$, define $\beta \in \mathrm{End}_{\mathbb{Q}}(T_1, F_1)$ by $\omega_1(\beta x, y) = \omega_2(\phi(x), \phi(y))$. The *Rosati involution* ($\beta \mapsto \beta'$) is the adjoint with respect to ω_1 and it fixes β since

$$\omega_1(\beta' x, y) = \omega_1(x, \beta y) = -\omega_1(\beta y, x) = -\omega_2(\phi(y), \phi(x)) = \omega_1(\beta x, y).$$

By Lemma C.1 there exists $\alpha \in \mathrm{End}_{\mathbb{Q}}(T_1, F_1)$ such that $\beta = \alpha'\alpha$ which gives

$$\omega_1(\alpha'\alpha x, y) = \omega_1(\alpha x, \alpha y) = \omega_2(\phi(x), \phi(y)).$$

Thus $\phi \circ \alpha^{-1} : (T_1 \otimes \bar{\mathbb{Q}}, F_1) \to (T_2 \otimes \bar{\mathbb{Q}}, F_2)$ is a $\bar{\mathbb{Q}}$-isogeny that preserves the polarizations.

Now suppose that real structures τ_1, τ_2 are provided. It is clear that (7) implies (6) and (4); also that (5)⇒(6)⇒(3). Now let us show (in the presence of τ_1, τ_2)) that (4)⇒(7). The involution $\tau_i \in \mathrm{GSp}(T_i, \omega_i)$ has multiplier -1. So by Lemma B.2 and Proposition B.4 there exists $\psi_i : T_i \otimes \mathbb{Q} \to \mathbb{Q}^{2n}$ which takes the symplectic form ω_i to the standard symplectic form ω_0, and which takes the involution τ_i to the standard involution τ_0. It therefore takes F_i to some $\gamma_i \in \mathrm{GSp}_{2n}(\mathbb{Q})$ which is q-inversive with respect to the standard involution τ_0.

By part (4), there is a $\bar{\mathbb{Q}}$ isogeny $\phi : (T_1, F_1, \lambda_1) \to (T_2, F_2, \lambda_2)$. This translates into an element $\theta \in \mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$ such that $\gamma_2 = \theta^{-1}\gamma_1\theta$.

By Proposition 5.5 there exists an element $\Psi \in \mathrm{GL}_n(\bar{\mathbb{Q}})$ such that $\gamma_2 = \Psi^{-1}\gamma_1\Psi$. In other words, Ψ corresponds to a $\bar{\mathbb{Q}}$-isogeny $(T_1, F_1, \lambda_1, \tau_1) \to (T_2, F_2, \lambda_2, \tau_2)$.

Now let us show that (6)⇒(5). Let us suppose that (T_1, F_1, τ_1) and (T_2, F_2, τ_2) are $\bar{\mathbb{Q}}$-isogenous. This implies that the characteristic polynomials $p_1(x)$ and $p_2(x)$ of F_1 and F_2 (respectively) are equal. Moreover, the ± 1 eigenspaces of τ_1 have the same dimension because $T \otimes \bar{\mathbb{Q}}_p$ decomposes as a direct sum of two subspaces that are exchanged by τ_1; see [Deligne 1969, §7]. Set $V_1 = T_1 \otimes \mathbb{Q}$ and $V_2 = T_2 \otimes \mathbb{Q}$ and denote these eigenspace decompositions as

$$V_1 \cong V_1^+ \oplus V_1^- \quad \text{and} \quad V_2 \cong V_2^+ \oplus V_2^-.$$

First let us consider the case that the characteristic polynomial $p_1(x)$ of F_1 is irreducible. In this case every nonzero vector in V_1 is a cyclic generator of V_1. Choose nonzero cyclic generators $v \in V_1^+$ and $w \in V_2^+$, and define $\psi : V \to V_2$ by

$$\psi(F_1^r v) = F_2^r w$$

for $1 \le r \le \dim(T)$. This mapping is well defined because F_1 and F_2 satisfy the same characteristic polynomial. Clearly, $\psi \circ F_1 = F_2 \circ \psi$. However we also claim that $\psi \circ \tau_1 = \tau_2 \circ \psi$. It suffices to check this on the cyclic basis which we do by induction. By construction we have that $\psi\tau_1 v = \tau_2\psi v = \tau_2 w$ so suppose we have proven that $\psi\tau_1 F_1^m v = \tau_2\psi F_1^m v = \tau_2 F_2^m w$ for all $m \le r-1$. Then

$$\begin{aligned}\psi\tau_1 F_1^r v &= \psi\tau_1 F_1 \tau_1^{-1}\tau_1 F_1^{r-1} v = q\psi F_1^{-1}\tau_1 F_1^{r-1} v = q F_2^{-1}\psi\tau_1 F_1^{r-1} v \\ &= q F_2^{-1}\tau_2\psi F_1^{r-1} v = \tau_2 F_2\psi F_1^{r-1} v = \tau_2\psi F_1^r v.\end{aligned}$$

Thus we have constructed a $\mathbb{Q}$ isogeny between these two real Deligne modules.

If the characteristic polynomial $p_1(x)$ is reducible then V_i $(i = 1, 2)$ may be decomposed as a direct sum of F_i-cyclic subspaces, each of which is preserved by the involution τ_i, thus reducing the problem to the case of irreducible characteristic polynomial. □

Proposition 6.2. *Associating the characteristic polynomial to each Deligne module induces a canonical one-to-one correspondence between*

(a) *the set of ordinary Weil q-polynomials $p(x) \in \mathbb{Z}[x]$ of degree $2n$ (see Appendix A),*

(b) *the set of $\mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$-conjugacy classes of semisimple elements $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ whose characteristic polynomial is an ordinary Weil q-polynomial,*

(c) *the set of $\mathbb{Q}$-isogeny classes of Deligne modules (T, F),*

(d) *the set of $\bar{\mathbb{Q}}$-isogeny classes of (Φ_ε-positively) polarized Deligne modules (T, F, λ),*

and a one-to-one correspondence between

(a′) *the set of ordinary real Weil q-polynomials (see Appendix A) of degree n,*

(b′) *the set of $\mathrm{GL}_n(\mathbb{Q})$-conjugacy classes of semisimple elements $A_0 \in \mathrm{GL}_n(\mathbb{Q})$ whose characteristic polynomial is $h(2x)$ where h is an ordinary real Weil q-polynomial,*

(c′) *the set of $\mathbb{Q}$-isogeny classes of Deligne modules (T, F, τ) with real structure,*

(d′) *the set of $\bar{\mathbb{Q}}$-isogeny classes of (Φ_ε-positively) polarized Deligne modules (T, F, ω, τ) with real structure.*

Proof. The correspondence (a)→(b) is given by Proposition A.5 (companion matrix for the symplectic group) while (b)→(a) associates to γ its characteristic polynomial. This correspondence is one-to-one since semisimple elements in $\mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$ are conjugate if and only if their characteristic polynomials are equal. Items (b) and (c) are identified by the Honda–Tate theorem [Tate 1971], but can also be seen directly. Given γ, one constructs a lattice $T \subset \mathbb{Q}^{2n}$ that is preserved by γ and by $q\gamma^{-1}$ by considering one cyclic subspace at a time (see Lemma 4.3) and taking T to be the lattice spanned by $\{\gamma^m v_0\}$ and by $\{(q\gamma)^m v_0\}$ where v_0 is a cyclic vector. Lemma 6.1 may be used to finish the proof that the correspondence is one-to-one. Items (c) and (d) correspond by Existence Lemma 3.10 (existence of a polarization) and Lemma 6.1.

The correspondence (a′)↔(b′) is standard. For the correspondence (a′)→(c′), each ordinary real Weil q-polynomial $h(x)$ is the real counterpart of an ordinary Weil q-polynomial $p(x)$ by Appendix A. It suffices to consider the case that $p(x)$ is irreducible. Let π be a root of $p(x)$ so that $K = \mathbb{Q}[\pi]$ is a CM field. Set $T = \mathcal{O}_K$ (the full ring of integers), let $F = \pi : T \to T$ be multiplication by π and let τ denote complex conjugation. Then τ preserves $\mathcal{O}_K$ and $\tau F \tau = qF^{-1}$ because $\pi\bar{\pi} = q$. Hence (T, F, π) is a Deligne module with real structure whose characteristic polynomial is $p(x)$. Lemma 6.1 says that this association (a′)→(c′) is one-to-one and onto. A mapping (c′)→(d′) is given by Existence Lemma 4.4 (existence of a polarization) and this mapping is one-to-one and onto by Lemma 6.1. □

In order to "count" the number of real Deligne modules it is necessary to describe them, up to isomorphism (rather than isogeny) in terms of algebraic groups as follows.

Lemma 6.3. *Let (T, F, ω, τ) be a rank $2n$ Deligne module with Φ_ε-positive polarization and real structure. Then it is isomorphic to one of the form*

$$(L, \gamma, \omega_0, \tau_0),$$

where $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ is q-inversive and its characteristic polynomial is an ordinary Weil q-polynomial; where $L \subset \mathbb{Q}^{2n}$ is a lattice that is preserved by γ, by $q\gamma^{-1}$ and by τ_0, and where the standard symplectic form ω_0 takes integer values on L. The group of self $\mathbb{Q}$-isogenies of (T, F, ω) (resp. (T, F, ω, τ)) is isomorphic to the centralizer $Z_\gamma(\mathbb{Q})$ in $\mathrm{GSp}_{2n}(\mathbb{Q})$ (resp. in $\mathrm{GL}_n^*(\mathbb{Q})$). Every element $\phi \in Z_\gamma(\mathbb{Q})$ has positive multiplier $c(\phi) > 0$.

If (T, F, ω, τ) is principally polarized then it is isomorphic to a principally polarized Deligne module of the form

$$(L_0, \gamma, \omega_0, \tau_1),$$

where $L_0 = \mathbb{Z}^{2n}$ is the standard lattice, where $\tau_1 \in \mathrm{GSp}_{2n}(\mathbb{Z})$ is an involution with multiplier -1, where $\gamma, q\gamma^{-1} \in \mathrm{GSp}_{2n}(\mathbb{Q}) \cap M_{2n\times 2n}(\mathbb{Z})$ preserve the integral lattice L_0 and where $\tau_1 \gamma \tau_1^{-1} = q\gamma^{-1}$.

Proof. By Lemma B.2 and Proposition B.4 there is a basis $\phi : T \otimes \mathbb{Q} \to \mathbb{Q}^{2n}$ of $T \otimes \mathbb{Q}$ so that ω becomes ω_0 and so that τ becomes τ_0. Take $L = \phi(T)$ and $\gamma = \phi F \phi^{-1}$. This induces an isomorphism $\mathbb{Q}[F] \cong \mathbb{Q}[\gamma]$ preserving the CM type Φ_ε on each, such that ω_0 is Φ_ε-positive. The centralizer statements are clear. If $\phi \in Z_\gamma(\mathbb{Q})$ and if $\iota \in \mathbb{Q}[\gamma]$ is a Φ_ε-totally positive imaginary element then

$$\begin{aligned} R(\phi(x), \phi(x)) &= \omega_0(\phi x, \iota \phi x) = \omega_0(\phi x . \phi \iota x) \\ &= c(\phi) R(x, x) > 0. \end{aligned}$$

If the original polarization ω of (T, F, τ) is principal then Lemma B.2 provides an isomorphism $(T, \omega) \cong (L_0, \omega_0)$ which takes F to some element γ and takes the involution τ to some involution τ_1, both of which preserve the lattice L_0. □

7. Finiteness

Throughout this section, all polarizations are considered to be Φ_ε-positive. As in Section 3, let $\mathbb{F}_q$ be a finite field of characteristic $p > 0$, fix $N \geq 1$ not divisible by p, and let $n \geq 1$. We refer to Section D.3 for the definition of a level N structure.

Theorem 7.1. *Assume $q, N \geq 1$ are coprime. There are finitely many isomorphism classes of principally (Φ_ε-positively) polarized Deligne modules of rank $2n$ over $\mathbb{F}_q$ with real structure and with principal level N structure.*

Proof. It follows from Proposition 6.2 that there are finitely many $\overline{\mathbb{Q}}$-isogeny classes of (Φ_ε-positively) polarized Deligne modules with real structure. Moreover, it is easy to see that each isomorphism class (of principally polarized Deligne modules with real structure) contains at most finitely many level N structures. So, for simplicity, we may omit the level structure, and it suffices to show that each $\overline{\mathbb{Q}}$-isogeny class (of Φ_ε-positively polarized Deligne modules with real structure) contains at most finitely many isomorphism classes of principally polarized modules. Therefore, let us fix a Φ_ε-positive principally polarized Deligne module with real structure, which by Lemma 6.3 may be taken to be of the form $(L_0, \gamma, \omega_0, \eta_0)$ where: $\eta_0 \in \mathrm{GSp}_{2n}(\mathbb{Z})$ is an involution with multiplier -1, where $\gamma_0 \in \mathrm{GSp}_{2n}(\mathbb{Q}) \cap M_{2n\times 2n}(\mathbb{Z})$ and its characteristic polynomial is an ordinary Weil q-polynomial, and where $\eta_0\gamma_0\eta_0^{-1} = q\gamma_0^{-1}$.

The group $G' = \mathrm{Sp}_{2n}$ acts on $V = M_{2n\times 2n} \times M_{2n\times 2n}$ by $g.(\gamma, \eta) = (g\gamma g^{-1}, g\eta g^{-1})$. Let $\Gamma = \mathrm{Sp}_{2n}(\mathbb{Z})$ be the arithmetic subgroup that preserves the lattice

$$L = M_{2n\times 2n}(\mathbb{Z}) \times M_{2n\times 2n}(\mathbb{Z})$$

of integral elements. It also preserves the set of pairs (γ, η) such that $\eta \in \mathrm{GSp}_{2n}(\mathbb{Z})$, $\eta^2 = I$, $\eta\gamma\eta^{-1} = q\gamma^{-1}$. Let $v_0 = (\gamma_0, \eta_0)$. We claim

(1) the orbit $G'_{\mathbb{C}}.v_0$ is closed in $V_{\mathbb{C}}$, and

(2) there is a natural injection from
   (a) the set of isomorphism classes of principally polarized abelian varieties with real structure within the $\overline{\mathbb{Q}}$-isogeny class of $(T_0, \gamma_0, \omega_0, \eta_0)$ to
   (b) the set of Γ-orbits in $L \cap G'_{\overline{\mathbb{Q}}}.v_0$.

Using claim (1) we may apply Borel's theorem[9] [1969, §9.11] and conclude that there are finitely many Γ orbits in $L \cap G'_{\overline{\mathbb{Q}}}.v_0$ which implies, by claim (2), that there are finitely many isomorphism classes.

Proof of claim (2). Consider a second principally polarized "real" Deligne module, $(L_0, \gamma_1, \omega_0, \eta_1)$, within the same $\overline{\mathbb{Q}}$-isogeny class. By Proposition 6.2, a $\overline{\mathbb{Q}}$-isogeny between these two Deligne modules is an element $X \in \mathrm{GSp}_{2n}(\overline{\mathbb{Q}})$ such that $\gamma_1 = X\gamma_0X^{-1}$ and $\eta_1 = X\eta_0X^{-1}$. In particular this means that the pair (γ_1, η_1) is in the orbit $\mathrm{GSp}_{2n}(\overline{\mathbb{Q}}).v_0$, which coincides with the orbit $G'_{\overline{\mathbb{Q}}}.v_0 = \mathrm{Sp}_{2n}(\overline{\mathbb{Q}}).v_0$. Moreover, such an isogeny X is an isomorphism (of principally polarized Deligne modules with real structure) if and only if X and X^{-1} preserve the lattice L_0 and the symplectic form ω_0, which is to say that $X \in \Gamma$.

[9] Let M be a reductive algebraic group defined over $\mathbb{Q}$ and let $\Gamma \subset M_{\mathbb{Q}}$ be an arithmetic subgroup. Let $M_{\mathbb{Q}} \to \mathrm{GL}(V_{\mathbb{Q}})$ be a rational representation of M on some finite-dimensional rational vector space. Let $L \subset V_{\mathbb{Q}}$ be a lattice that is stable under Γ. Let $v_0 \in V$ and suppose that the orbit $M_{\mathbb{C}}.v_0$ is closed in $V_{\mathbb{C}} = V_{\mathbb{Q}} \otimes \mathbb{C}$. Then $L \cap M_{\mathbb{C}}.v_0$ consists of a finite number of orbits of Γ.

We remark that the mapping from (2a) to (2b) above is not necessarily surjective for the following reason. The element γ_0 is (Φ_ε-)*viable* (see Section 3.11), that is, it satisfies the "positivity" condition $(*)$ of Section 3.7, because it comes from a polarized abelian variety. However, if $(\gamma, \eta) \in L \cap G'_{\bar{\mathbb{Q}}}.v_0$ is arbitrary then γ may fail to be Φ_ε-viable.

Proof of claim (*1*). Since γ_0 and η_0 are both semisimple, the conjugacy class

$$(G'_{\mathbb{C}}.\gamma_0) \times (G'_{\mathbb{C}}.\eta_0) \subset M_{2n\times 2n}(\mathbb{C}) \times M_{2n\times 2n}(\mathbb{C})$$

is closed [Humphreys 1975, §18.2]. We claim that the orbit $G'_{\mathbb{C}}.v_0$ coincides with the closed subset

$$S = \left\{(\gamma, \tau) \in (G'_{\mathbb{C}}.\gamma_0) \times (G'_{\mathbb{C}}.\tau_0) |\ \tau\gamma\tau^{-1} = q\gamma^{-1}\right\}.$$

Clearly, $G'_{\mathbb{C}}.v_0 \subset S$. If $(\gamma, \eta) \in (G'_{\mathbb{C}}.\gamma_0) \times (G'_{\mathbb{C}}.\eta_0)$ lies in the subset S then by Proposition B.4, conjugating by an element of $G'_{\mathbb{C}}$ if necessary, we may arrange that $\eta = \tau_0$ is the standard involution. Consequently, $\tau_0\gamma\tau_0^{-1} = q\gamma^{-1}$, which is to say that γ is q-inversive. By assumption, it is also $G'_{\mathbb{C}}$-conjugate to γ_0. According to Proposition 5.5, $G'_{\mathbb{C}}$-conjugacy of q-inversive elements coincides with $\delta(\mathrm{GL}_n(\mathbb{C}))$-conjugacy. Thus there exists $g \in \delta(\mathrm{GL}_n(\mathbb{C}))$ such that $(g\gamma g^{-1}, g\tau_0 g^{-1}) = (\gamma_0, \tau_0)$. In summary, the element (γ, η) lies in the $G'_{\mathbb{C}}$-orbit of (γ_0, τ_0). This concludes the proof of Theorem 7.1. □

7.2. ***The case n = 1.*** Fix $q = p^m$ and let $\mathbb{F}_q$ denote the finite field with q elements. According to Proposition 6.2, the set of $\bar{\mathbb{Q}}$-isogeny classes of Deligne modules (T, F) of rank 2 over $\mathbb{F}_q$ is determined by a quadratic ordinary Weil q-number π, which we now fix. This means that π satisfies an equation

$$\pi^2 + B\pi + q = 0$$

where $p \nmid B$. Let $D = B^2 - 4q$. Then $D \equiv 0, 1 \pmod 4$ and $-4q < D < 0$. The pair $\{\pi, \bar{\pi}\}$ determines D and vice versa.

Isomorphism classes of polarized Deligne modules with real structure fall into orbits that are identified by certain cohomology classes as described in Proposition D.7 or, equivalently, identified by integral conjugacy classes of involutions as described in Proposition D.2. For $n = 1$ there are two involutions (see Lemma B.5) to consider, namely

$$\tau_0 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \tau_1 = \begin{pmatrix} -1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Proposition 7.3. *Over the finite field $\mathbb{F}_q$, the number of* (*real isomorphism classes of*) *principally polarized Deligne modules (T, F, λ, η) with real structure and*

rank 2, *such that the eigenvalues of* F *are* $\{\pi, \bar{\pi}\}$, *which correspond to the cohomology class of the standard involution* τ_0 *is*

$$\begin{cases} \sigma_0(-D/4) & \text{if } D \equiv 0 \pmod 4, \\ 0 & \text{otherwise}, \end{cases}$$

where $\sigma_0(m)$ *denotes the number of positive divisors of* $m > 0$. *The number of isomorphism classes which correspond to the cohomology class of* τ_1 *is*

$$\begin{cases} \sigma_0(-D) & \text{if } D \equiv 1 \pmod 4, \\ \sigma_0'(-D/4) & \text{if } D \equiv 0 \pmod 4, \end{cases}$$

where $\sigma_0'(m)$ *denotes the number of ordered factorizations* $m = uv$ *such that* $u, v > 0$ *have the same parity.*

Proof. According to Proposition 6.2 the isomorphism classes of principally polarized Deligne modules with real structures correspond to q-inversive pairs (γ, η) where the eigenvalues of $\gamma \in \mathrm{GL}_2(\mathbb{Z})$ are π and $\bar{\pi}$. For the involution τ_0, the pair (γ, τ_0) is q-inversive if $\gamma = \left(\begin{smallmatrix} a & b \\ c & a \end{smallmatrix}\right)$ and $\det(\gamma) = q$. This implies that $a = -B/2$, so B is even and $D \equiv 0 \pmod 4$. Then $bc = a^2 - q = D/4$ has a unique solution for every (signed) divisor b of $D/4$. Half of these will be viable (see Section 3.11) so the number of solutions is equal to the number of positive divisors of $-D/4$.

For the involution τ_1, the pair (γ, τ_1) is q-inversive if $\gamma = \left(\begin{smallmatrix} a & b \\ c & a-b \end{smallmatrix}\right)$. This implies that $D = B^2 - 4q = b(b + 4c)$. Let us first consider the case that b is odd or equivalently, that $D \equiv 1 \pmod 4$. For every divisor $b \mid D$ we can solve for an integer value of c so we conclude that the number of viable solutions in this case is equal to $\sigma_0(-D)$. Next, suppose that b is even, say, $b = 2b'$. Then D is divisible by 4, say, $D = 4D'$ and $D' = b'(b' + 2c)$ is an ordered factorization of D' with factors of the same parity. So in this case the number of viable solutions is $\sigma_0'(-D/4)$. □

7.4. It follows that the total number of real isomorphism classes over $\mathbb{F}_q$, $q = p^m$, corresponding to the trivial cohomology class, is $N = 2\sum_{1 \le a \le q-1} \sigma_0(q - a^2)$, a number whose asymptotics was determined by Ingham [1927] and Hooley [1958],

$$N \sim \begin{cases} \frac{6}{\pi^2}(\sqrt{q}(\log(q))^2 + 3\log 2 \log q), & m \text{ even}, \\ C(p)\sqrt{q}\log q, & m \text{ odd}. \end{cases}.$$

7.5. For any totally positive imaginary integer $\alpha \in L = \mathbb{Q}(\pi)$, the bilinear form $\omega(x, y) = \mathrm{Trace}_{L/\mathbb{Q}}(\alpha x \bar{y})$ is symplectic. If $\Lambda \subset L$ is a lattice then α may be chosen so that the form ω takes integer values on Λ. Modifying Λ by a homothety if necessary, it can also be arranged that ω is a principal polarization, hence (Λ, π, ω) is a principally polarized Deligne module. If complex conjugation on $L = \mathbb{Q}(\pi)$ preserves Λ then it defines a real structure on this Deligne module.

Proposition 7.6. *The set of isomorphism classes of Φ_ε-positive principally polarized Deligne modules (of rank 2) with real structure and with eigenvalues $\{\pi, \bar{\pi}\}$ may be identified with the set of homothety classes of lattices $\Lambda \subset \mathbb{Q}(\pi)$ that are preserved by complex conjugation and by multiplication by π.*

Proof. The most natural proof, which involves considerable checking, provides a map back from lattices Λ to Deligne modules: Deligne's CM type determines an isomorphism $\Phi : \mathbb{Q}(\pi) \otimes \mathbb{R} \to \mathbb{C}$. Then realize the elliptic curve $\mathbb{C}/\Phi(\Lambda)$ as the complex points of the canonical lift of an ordinary elliptic curve over $\mathbb{F}_q$ whose associated Deligne module is (Λ, π). Then check that complex conjugation is compatible with these constructions.

A simpler but less illuminating proof is simply to count the number of homothety classes of lattices and to see this number coincides with that in Proposition 7.3. □

8. $\mathbb{Q}$-isogeny classes within a $\bar{\mathbb{Q}}$ isogeny class

8.1. Let us fix a (Φ_ε-positively) polarized Deligne module with real structure, $(L_1, \gamma_1, \omega_0, \tau_0)$ where

$$\gamma_1 = \begin{pmatrix} A_1 & B_1 \\ C_1 & {}^tA_1 \end{pmatrix}$$

is q-inversive, ω_0 is the standard symplectic form, τ_0 is the standard involution, and $L_1 \subset \mathbb{Q}^{2n}$ is a lattice preserved by τ_0, by γ_1 and by $q\gamma_1^{-1}$, on which ω_0 takes integer values; see Lemma 6.3 . Let $Z_{\mathrm{GL}_n(\mathbb{Q})}(A_1)$ denote the set of elements in $\mathrm{GL}_n(\mathbb{Q})$ that commute with A_1.

Proposition 8.2. *The association $C \mapsto \gamma = \left(\begin{smallmatrix} A_1 & B \\ C & {}^tA_1 \end{smallmatrix}\right)$, where $B = (A_1 - qI)C^{-1}$, determines a one-to-one correspondence between*

(1) *the set of elements $C \in \mathrm{GL}_n(\mathbb{Q})$, one from each $Z_{\mathrm{GL}_n(\mathbb{Q})}(A_1)$-congruence class of symmetric matrices such that ${}^tA_1 C = C A_1$ and $\mathrm{sig}(A_1; C) = \mathrm{sig}(A_1; C_1)$,*

(2) *the set of $\mathbb{Q}$ isogeny classes of real (Φ_ε-positively) polarized Deligne modules (T, F, λ, τ) within the $\bar{\mathbb{Q}}$ isogeny class of $(L_1, \gamma_1, \omega_0, \tau_0)$,*

(3) *the set of $\mathrm{GL}_n^*(\mathbb{Q})$-conjugacy classes of elements $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ such that γ, γ_1 are conjugate by some element in $\mathrm{GL}_n^*(\mathbb{R}) \subset \mathrm{Sp}_{2n}(\mathbb{R})$,*

(4) *the set elements of $\ker(H^1(\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}), I_1) \to H^1(\mathrm{Gal}(\mathbb{C}/\mathbb{R}), I_1),)$*

where I_1 denotes the group of self isogenies of $(L_1, \gamma_1, \omega_0, \tau_0)$; that is,

$$I_1 = Z_{\mathrm{GL}_n^*}(\gamma_1) \cong Z_{\mathrm{GL}_n}(A_1) \cap \mathrm{GO}(C_1), \tag{8.2.1}$$

where

$$\mathrm{GO}(C_1) = \{X \in \mathrm{GL}_n \mid {}^tXC_1X = \mu C_1 \text{ (there exists } \mu \neq 0)\}$$

denotes the general orthogonal group defined by the symmetric matrix C_1.

Proof. To describe the correspondence (1)→(3), given C, set $\gamma = \left(\begin{smallmatrix} A_1 & B \\ C & {}^tA_1 \end{smallmatrix}\right)$ where $B = (A^2 - qI)C^{-1}$. Since $\mathrm{sig}(A_1; C) = \mathrm{sig}(A_1; C_1)$, Proposition 5.5 implies that γ and γ_1 are conjugate by an element of $\delta(\mathrm{GL}_n(\mathbb{R})) \subset \mathrm{GL}_n^*(\mathbb{R})$.

Conversely, let $\gamma = \left(\begin{smallmatrix} A & B \\ C & {}^tA \end{smallmatrix}\right) \in \mathrm{GSp}_{2n}(\mathbb{Q})$ be q-inversive and $\mathrm{GL}_n^*(\mathbb{R})$-conjugate to γ_1. Then A, A_1 are conjugate by an element of $\mathrm{GL}_n(\mathbb{R})$ so they are also conjugate by some element $Y \in \mathrm{GL}_n(\mathbb{Q})$. Replacing γ with $\delta(Y)\gamma\delta(Y)^{-1}$ we may therefore assume that $A = A_1$. Proposition 5.5 then says that $\mathrm{sig}(A_1; C) = \mathrm{sig}(A_1; C_1)$. So we have a one-to-one correspondence (1)↔(3).

For (2)→(3), let $(L, \gamma, \omega_0, \tau_0)$ be a Φ_ε-positively polarized Deligne module with real structure that is $\bar{\mathbb{Q}}$-isogenous to $(L_1, \gamma_1, \omega_0, \tau_0)$. Then γ_1, γ_2 are Φ_ε-viable so by Proposition 3.12 they are also $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugate. Proposition 5.5 says they are $\mathrm{GL}_n^*(\mathbb{R})$-conjugate. A choice of $\bar{\mathbb{Q}}$ isogeny $\phi : (L, \gamma, \omega_0, \tau_0) \to (L_1, \gamma_1, \omega_0, \tau_0)$ is an element $X \in \mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$ such that $\gamma = X\gamma_1 X^{-1}$ and $\tau_0 X \tau_0^{-1} = X$, hence $X \in \mathrm{GL}_n^*(\bar{\mathbb{Q}})$. The isogeny ϕ is a $\mathbb{Q}$-isogeny if and only if $X \in \mathrm{GL}_n^*(\mathbb{Q})$.

For (3)→(2), start with the basepoint $(L_1, \gamma_1, \omega_0, \tau_0)$ and choose any element $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ that is $\mathrm{GL}_n^*(\mathbb{R})$-conjugate to γ_1. Then

$$\gamma = t\gamma_1 t^{-1} = h\gamma_1 h^{-1} \tag{8.2.2}$$

for some $t \in \mathrm{GL}_{2n}(\mathbb{Q})$ and some $h \in \mathrm{GL}_n^*(\mathbb{R})$. The set

$$L' := (tL_1) \cap (\tau_0 t L_1) \subset \mathbb{Q}^{2n}$$

is a lattice, so there exists an integer m such that ω_0 takes integer values on $L := mL'$. Then $(L = mL', \gamma, \omega_0, \tau_0)$ is a polarized Deligne module with real structure in the $\bar{\mathbb{Q}}$-isogeny class of $(L_1, \gamma_2, \omega_0, \tau_0)$. The lattice L is preserved by τ_0, by γ and by $q\gamma^{-1}$ from (8.2.2). The element γ is Φ_ε-viable by construction so the symplectic form ω_0 is a polarization on the Deligne module (L, γ).

For (3)↔(4), the set $H^1(\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}), I_1)$ indexes the $\mathrm{GL}_n^*(\mathbb{Q})$-conjugacy classes of elements γ within the $\mathrm{GL}_n^*(\bar{\mathbb{Q}})$-conjugacy class of γ_1. Such a class becomes trivial in $H^1(\mathrm{Gal}(\mathbb{C}/\mathbb{R}), I_1)$ if γ is $\mathrm{GL}_n^*(\mathbb{R})$-conjugate to γ_1. The isomorphism of (8.2.1) follows immediately from (5.6.2). □

There may be infinitely many $\mathbb{Q}$-isogeny classes of polarized Deligne modules with real structure within a given $\bar{\mathbb{Q}}$-isogeny class, but from Theorem 7.1, only finitely many of these $\mathbb{Q}$-isogeny classes contain principally polarized modules.

8.3. Let $Z(\gamma_1)$ denote the centralizer of γ_1 in GSp_{2n}. Removing the real structure from the proof of Proposition 8.2 gives a one-to-one correspondence between (a) the set of $\mathbb{Q}$-isogeny classes of Φ_ε-positively polarized Deligne modules within the $\bar{\mathbb{Q}}$-isogeny class of $(L_1, \gamma_1, \omega_0)$, (b) the set of $\mathrm{Sp}_{2n}(\mathbb{Q})$-conjugacy classes of elements $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ that are $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugate to γ_1, and (c) elements of

$$\ker(H^1(\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}), Z(\gamma_1)) \to H^1(\mathrm{Gal}(\mathbb{C}/\mathbb{R}), Z(\gamma_1)).$$

9. Isomorphism classes within a $\mathbb{Q}$-isogeny class

9.1. ***The category $\mathcal{P}_N$.*** In this section and in all subsequent sections we fix $N \geq 3$, not divisible by p. Fix $n \geq 1$. Throughout this section we fix a Φ_ε-positively polarized Deligne module (over $\mathbb{F}_q$, of rank $2n$) with real structure, which (by Lemma 6.3) we may assume to be of the form $(T_0, \gamma, \omega_0, \tau_0)$ where $T_0 \subset \mathbb{Q}^{2n}$ is a lattice, $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ is a semisimple element whose characteristic polynomial is an ordinary Weil q-polynomial, and where ω_0 is the standard symplectic form and τ_0 is the standard involution.

Following the method of [Kottwitz 1990] we consider the category $\mathcal{P}_N = \mathcal{P}_N(T_0, \gamma, \omega_0, \tau_0)$ (possibly empty) for which an object $(T, F, \omega, \beta, \tau, \phi)$ consists of a (Φ_ε positive) *principally* polarized Deligne module $(T, F, \omega, \beta, \tau)$ with real structure τ and principal level N structure β (see Definition 4.1), and where $\phi : (T, F, \omega, \tau) \to (T_0, \gamma, \omega_0, \tau_0)$ is a $\mathbb{Q}$-isogeny of polarized Deligne modules with real structure, meaning $\phi : T \otimes \mathbb{Q} \xrightarrow{\sim} \mathbb{Q}^{2n}$, $\phi F = \gamma\phi$, $\phi^*(\omega_0) = c\omega$ for some $c \in \mathbb{Q}^\times$, and $\phi\tau = \tau_0\phi$. A morphism

$$\psi : (T, F, \omega, \beta, \tau, \phi) \to (T', F', \omega', \beta', \tau', \phi')$$

is a group homomorphism $\psi : T \hookrightarrow T'$ such that $\phi = \phi'\psi$ (hence $\psi F = F'\psi$), $\omega = \psi^*(\omega')$, $\beta = \beta' \circ \psi$, and $\psi\tau = \tau'\psi$.

Let X denote the set of isomorphism classes in this category. We obtain a natural one-to-one correspondence between the set of isomorphism classes of principally polarized Deligne modules with level N structure and real structure within the $\mathbb{Q}$-isogeny class of $(T_0, \gamma, \omega_0, \tau_0)$, and the quotient

$$I_{\mathbb{Q}} \backslash X, \tag{9.1.1}$$

where $I_{\mathbb{Q}} = I_{\mathbb{Q}}(T_0, \gamma, \omega_0, \tau_0)$ denotes the group of self $\mathbb{Q}$-isogenies of $(T_0, \gamma, \omega_0, \tau_0)$.

9.2. ***Category of lattices.*** See Section 1.7 for the notation $\mathbb{A}_f^p$, $\widehat{\mathbb{Z}}^p$, $\widehat{K}_N$, $\widehat{K}_N^0$, K_p. Denote by $\mathcal{L}_N = \mathcal{L}_N(\mathbb{Q}^{2n}, \gamma, \omega_0, \tau_0)$ the category for which an object is a pair (L, α) where $L \subset T_0 \otimes \mathbb{Q}$ is a lattice that is symplectic (up to homothety), is preserved by γ, by $q\gamma^{-1}$ and by τ_0, and $\alpha : L/NL \to (\mathbb{Z}/N\mathbb{Z})^{2n}$ is a compatible level structure (that is, $\bar{\tau}_0\alpha = \alpha\bar{\tau}_0$ and $\alpha\gamma = \alpha$), and there exists $c \in \mathbb{Q}^\times$ so that $L^\vee = cL$ and $\alpha_*(c\omega_0) = \overline{\omega}_0$.

A morphism $(L, \alpha) \to (L', \alpha')$ is an inclusion $L \subset L'$ such that $\alpha' \mid (L/NL) = \alpha$. (Since $L \to L'$ is an inclusion, it also commutes with γ and τ_0, and it preserves the symplectic form ω_0.) In this category every isomorphism class contains a unique object.

9.3. ***Adèlic lattices.*** Given $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ as above, let $\widehat{\mathcal{L}}_N = \widehat{\mathcal{L}}_N(\mathbb{A}_f^{2n}, \gamma, \omega_0, \tau_0)$ be the category for which an object is a pair $(\widehat{L}, \alpha)$ consisting of a lattice $\widehat{L} \subset \mathbb{A}_f^{2n}$

(see Sections D.3, D.4, D.5) that is symplectic (up to homothety) and is preserved by γ, by $q\gamma^{-1}$ and by τ_0, and a compatible level N structure α, that is:

$$\tau_0\widehat{L} = \widehat{L}, \quad \alpha\bar{\tau}_0 = \bar{\tau}_0\alpha, \tag{9.3.0}$$

$$\gamma\widehat{L} \subset \widehat{L}, \quad q\gamma^{-1}\widehat{L} \subset \widehat{L}, \quad \alpha \circ \gamma = \alpha, \tag{9.3.1}$$

$$\widehat{L}^\vee = c\widehat{L} \quad \text{and} \quad \alpha_*(c\omega_0) = \overline{\omega}_0, \quad (\text{there exists } c \in \mathbb{Q}^\times). \tag{9.3.2}$$

A morphism in $\widehat{\mathcal{L}}_N$ is an inclusion $\widehat{L} \subset \widehat{M}$ that is compatible with the level structures. As in [Kottwitz 1990] we have the following:

Proposition 9.4. *The association*

$$(T, F, \omega, \beta, \tau, \phi) \mapsto (L = \phi(T), \alpha = \beta \circ \phi^{-1}) \mapsto \left(\widehat{L} = \prod_v L \otimes \mathbb{Z}_v, \alpha\right)$$

determines equivalences of categories $\mathcal{P}_N \xrightarrow{\sim} \mathcal{L}_N \xrightarrow{\sim} \widehat{\mathcal{L}}_N$.

Proof. Given $(T, F, \omega, \beta, \phi)$, let $L = \phi(T)$ and $\alpha = \beta\phi^{-1}$. Then $\gamma L = \gamma\phi(T) = \phi(FT) \subset \phi(T) = L$ and similarly $q\gamma^{-1}L \subset L$. Since ω is a principal polarization we obtain

$$T = T^\vee = \{u \in T \otimes \mathbb{Q} \mid \omega(u, v) \in \mathbb{Z} \text{ for all } v \in T\}.$$

Since ϕ is a $\mathbb{Q}$-isogeny with multiplier $c \in \mathbb{Q}^\times$ we have that $\omega_0(\phi(x), \phi(y)) = c\omega(x, y)$ for all $x, y \in T \otimes \mathbb{Q}$ so

$$\begin{aligned} L^\vee &= \{u \in L \otimes \mathbb{Q} \mid \omega_0(u, v) \in \mathbb{Z} \text{ for all } v \in L\} \\ &= (\phi(T))^\vee = c\phi(T^\vee) = c\phi(T) = cL. \end{aligned}$$

This implies that $c\omega_0$ is integral-valued on L and hence

$$\alpha_*(c\omega_0) = \beta_*\phi^*(c\omega_0) = \beta_*(\omega) = \overline{\omega}_0.$$

Hence the pair (L, α) is an object in $\mathcal{L}_N(\mathbb{Q}^{2n}, \gamma, \omega_0, \tau_0)$. If

$$\psi : (T, F, \omega, \phi) \to (T', F', \omega', \phi')$$

is a morphism in $\mathcal{P}_N(T_0, \gamma, \omega_0, \tau_0)$ then $\psi(T) \subset T'$ so $L = \phi(T) \subset L' = \phi(T')$ is a morphism in $\mathcal{L}_N(\mathbb{Q}^{2n}, \gamma, \omega_0, \tau_0)$.

Conversely, given an object (L, α) in $\mathcal{L}_N$, that is, a lattice $L \subset \mathbb{Q}^{2n}$ preserved by γ and $q\gamma^{-1}$ such that $L^\vee = cL$, and a principal level structure $\alpha : L/NL \to (\mathbb{Z}/N\mathbb{Z})^{2n}$ such that $\alpha_*(c\omega_0) = \overline{\omega}_0$, we obtain an object in $\mathcal{P}_N$,

$$(T = L,\ F = \gamma \mid L,\ \omega = c\omega_0,\ \beta = \alpha,\ \phi = \mathrm{id})$$

such that $T^\vee = \frac{1}{c}L^\vee = L = T$ and such that $\beta_*(\omega) = \alpha_*(c\omega_0) = \overline{\omega}_0$. It follows that $\mathcal{P}_N \to \mathcal{L}_N$ is an equivalence of categories. Finally, $\mathcal{L}_N \to \widehat{\mathcal{L}}_N$ is an equivalence of categories by Lemma D.5. □

9.5. ***Lattices at p.*** Let $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ be a semisimple element whose characteristic polynomial is an ordinary Weil q-polynomial. It induces a decomposition $\mathbb{Q}_p^{2n} \cong W' \oplus W''$ that is preserved by γ but exchanged by τ_0, where the eigenvalues of $\gamma \mid W'$ are p-adic units and the eigenvalues of $\gamma \mid W''$ are nonunits [Deligne 1969, §7]. Define

$$\alpha_q = \alpha_{\gamma,q} = I' \oplus qI'' \tag{9.5.1}$$

to be the identity on W' and multiplication by q on W''. The following lemma, which is implicit in [Kottwitz 1990] will be used in Proposition 9.8 to count the number of lattices in $\widehat{\mathcal{L}}_N$.

Lemma 9.6. *Let $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ be a semisimple element with multiplier equal to q. Let $L_{0,p} = \mathbb{Z}_p^{2n}$ be the standard lattice in $\mathbb{Q}_p^{2n}$. Let $g \in \mathrm{GSp}_{2n}(\mathbb{Q}_p)$ and let c denote its multiplier. Let $L_p = g(L_{0,p})$. Then $L_p^\vee = c^{-1}L_p$ and the following statements are equivalent*:

(a) *The lattice L_p is preserved by γ and by $q\gamma^{-1}$.*

(b) *The lattice L_p satisfies $qL_p \subset \gamma L_p \subset L_p$.*

(c) *$g^{-1}\gamma g \in K_p A_q K_p$ where $K_p = \mathrm{GSp}_{2n}(\mathbb{Z}_p)$ and $A_q = \left(\begin{smallmatrix} I & 0 \\ 0 & qI \end{smallmatrix}\right)$.*

If the characteristic polynomial of γ is an ordinary Weil q-polynomial then conditions (a), (b), (c) above are also equivalent to

(d) *$g^{-1}\alpha_q^{-1}\gamma g \in K_p$.*

Proof. Clearly (a) and (b) are equivalent, and also to $qL_p \subset q\gamma^{-1}L_p \subset L_p$. Hence

(b′) $L_p/\gamma L_p \cong \gamma^{-1}L_p/L_p \subset L_p/qL_p \cong (\mathbb{Z}/q\mathbb{Z})^{2n}$.

We now show that (b)⇒(c). Since $\det(\gamma)^2 = q^{2n}$ we know that $|\det(\gamma)| = |L_p/\gamma L_p| = q^n$. Condition (b) implies that $L_p/\gamma L_p$ consists of elements that are killed by multiplication by q. Condition (b′) implies that $L_p/\gamma L_p$ is free over $\mathbb{Z}/q\mathbb{Z}$. Therefore,

$$L_{0,p}/(g^{-1}\gamma g)L_{0,p} \cong L_p/\gamma L_p \cong (\mathbb{Z}/q\mathbb{Z})^n. \tag{9.6.1}$$

By the theory of Smith normal form for the symplectic group (see [Spence 1972] or [Andrianov 1987, Lemma 3.3.6]), we may write $g^{-1}\gamma g = uDv$ where $u, v \in K_p$ and $D = \mathrm{diag}(p^{r_1}, p^{r_2}, \ldots, p^{r_{2n}})$ where $r_1 \le r_2 \le \cdots \le r_{2n}$. This, together with (9.6.1) implies that $r_1 = \cdots = r_n = 0$ and $r_{n+1} = \cdots = r_{2n} = a$; that is, $D = A_q$. This proves that (b) implies (c).

Now we show (c) implies (a). Since $K_p A_q K_p \subset M_{2n\times 2n}(\mathbb{Z}_p)$, condition (c) implies $\gamma g L_{0,p} \subset g L_{0,p}$. Taking the inverse of condition (c) and multiplying by q gives

$$qg^{-1}\gamma^{-1}g \in K_p q A_q^{-1} K_p \subset M_{2n\times 2n}(\mathbb{Z}_p),$$

which implies that $q\gamma^{-1}L_p \subset L_p$.

Finally, if the characteristic polynomial of γ is an ordinary Weil q-polynomial then (by [Deligne 1969]) the lattice L_p decomposes into γ-invariant sublattices, $L_p = L'_p \oplus L''_p$ such that $\gamma \mid L'_p$ is invertible and $\gamma \mid L''_p$ is divisible by q, or $\gamma L'_p = L'_p$ and $\gamma L''_p \subset qL''_p$ which, in light of (d) implies that $\gamma L''_p = qL''_p$. In summary, $\alpha_q^{-1}\gamma L_p = L_p$, which is equivalent to (d). □

9.7. ***Counting real lattices.*** As explained in Section 9.1, we wish to count the number of isomorphism classes of (Φ_ε-positive) principally polarized Deligne modules with level N structure and with real structure that are $\mathbb{Q}$-isogenous to the polarized Deligne module $(T_0, \gamma, \omega_0, \tau_0)$ that was fixed in Section 9.1. By (9.1.1) and Proposition 9.4, this number is

$$|S(\mathbb{Q})\backslash X|,$$

where X denotes the set of objects $(\widehat{L}, \alpha)$ in the category $\widehat{\mathcal{L}}_N(\mathbb{A}_f^{2n}, \gamma, \omega_0, \tau_0)$ of Section 9.3 and where $S(\mathbb{Q})$ denotes the group of (involution-preserving) $\mathbb{Q}$-self isogenies of $(T_0, \gamma, \omega_0, \tau_0)$. It may be identified with the centralizer

$$S_\gamma(\mathbb{Q}) = \{x \in \mathrm{GL}_n^*(\mathbb{Q}) \mid \gamma x = x\gamma\}.$$

(Note that $\gamma \notin \mathrm{GL}_n^*(\mathbb{Q})$.) Following Proposition D.7, the $\mathrm{GL}_n^*(\mathbb{A}_f)$-orbit containing a given object $(\widehat{L}, \alpha)$ is determined by its cohomology class

$$[\widehat{L}, \alpha] \in H^1 = H^1(\langle\tau_0\rangle, \widehat{K}_N^0)$$

of (D.7.1). For simplicity, for the moment we assume that N is even: this implies that the contributions from different cohomology classes are independent of the cohomology class, as explained in the following paragraph.

Fix such a class $[t] \in H^1$, corresponding to some element $t \in \widehat{K}_N^0$ with $t\tilde{t} = 1$. Let

$$X_{[t]} = \{(\widehat{L}, \alpha) \in X \mid [(\widehat{L}, \alpha)] = [t] \in H^1\}$$

denote the set of objects $(\widehat{L}, \alpha)$ whose associated cohomology class is $[t]$. We wish to count the number of elements in the set $S_\gamma(\mathbb{Q})\backslash X_t$. Since N is even, the cohomology class $[t]$ vanishes in the cohomology of $\mathrm{Sp}_{2n}(\widehat{\mathbb{Z}})$, by Proposition D.9. This means that $t = g^{-1}\tilde{g}$ for some $g \in \mathrm{Sp}_{2n}(\widehat{\mathbb{Z}})$.

Let $\widehat{L}_0 = \widehat{Z}^{2n}$ and $\alpha_0 : \widehat{L}/N\widehat{L} \to (\mathbb{Z}/N\mathbb{Z})^{2n}$ denote the standard lattice and the standard level N structure. Then $(g\widehat{L}_0, \alpha_0 \circ g^{-1}) \in \mathcal{R}_N$ is a lattice with real structure and level N structure, whose cohomology class equals $[t] \in H^1$. Its isotropy group under the action of $\mathrm{GL}_n^*(\mathbb{A}_f)$ is the principal congruence subgroup

$$\begin{aligned}\widehat{\Gamma}_N &= \mathrm{GL}_n^*(\mathbb{A}_f) \cap g\widehat{K}_N g^{-1} \\ &= \mathrm{GL}_n^*(\mathbb{A}_f) \cap \widehat{K}_N\end{aligned}$$

(since $\widehat{K}_N$ is a normal subgroup of $\mathrm{Sp}_{2n}(\widehat{\mathbb{Z}})$) and is independent of the class $[t]$.

Hence

$$X_{[t]} \cong \widehat{\Gamma}_N \backslash \mathrm{GL}_n^*(\mathbb{A}_f)$$

is a finite-adèlic analog of the space $X_{\mathbb{C}}$ described in Section 1.1. Choose the Haar measure on $\mathrm{GL}_n^*(\mathbb{A}_f)$ that gives measure one to the group $\widehat{\Gamma}_N$.

With $\widehat{K}_N = \widehat{K}_N^p K_p$ (see Section 1.7), define χ^p to be the characteristic function on $\mathrm{GSp}_{2n}(\mathbb{A}_f^p)$ of $\widehat{K}_N^p$ and define χ_p to be the characteristic function on $\mathrm{GSp}_{2n}(\mathbb{Q}_p)$ of K_p. Let $H = \mathrm{GL}_n^*$.

Proposition 9.8. *Suppose that $N \geq 3$ is even and $p \nmid N$. Then*

$$|S_\gamma(\mathbb{Q})\backslash X_t| = \mathrm{vol}(S_\gamma(\mathbb{Q})\backslash S_\gamma(\mathbb{A}_f)) \cdot I_\gamma^p \cdot I_{\gamma,p},$$

where

$$I_\gamma^p = \int_{S_\gamma(\mathbb{A}_f^p)\backslash H(\mathbb{A}_f^p)} \chi^p(x^{-1}\gamma x)\, dx \tag{9.8.1}$$

and

$$I_{\gamma,p} = \int_{S_\gamma(\mathbb{Q}_p)\backslash H(\mathbb{Q}_p)} \chi_p(x^{-1}\alpha_q^{-1}\gamma x)\, dx.$$

Here, $\alpha_q = \alpha_{\gamma,q}$ is defined in (9.5.1).

Proof. By Proposition D.7 each $(\widehat{L}, \alpha) \in X_{[t]}$ has the form $xg.(\widehat{L}_0, \alpha_0)$ for some

$$x = (x^p, x_p) \in \mathrm{GL}_n^*(\mathbb{A}_f^p) \times \mathrm{GL}_n^*(\mathbb{Q}_p) = \mathrm{GL}_n^*(\mathbb{A}_f),$$

where $t = g^{-1}\tilde{g}$ as above, with $g \in \mathrm{Sp}_{2n}(\widehat{\mathbb{Z}})$. Write $\widehat{L} = L^p \times L_p$ for its component away from p and component at p, respectively, and similarly for $g = g^p g_p$. The conditions (9.3.1) give $\gamma x^p g^p L_0^p = x^p g^p L_0^p$. Hence

$$(g^p)^{-1}(x^p)^{-1}\gamma x^p g^p \in \widehat{K}_N^p,$$

which is normal in K^p so, equivalently, $\chi^p((x^p)^{-1}\gamma x^p) = 1$. Similarly, by Lemma 9.6,

$$g_p^{-1}x_p^{-1}\alpha_q^{-1}\gamma x_p g_p \in K_p \quad \text{or} \quad \chi_p(x_p^{-1}\alpha_q^{-1}\gamma x_p) = 1.$$

In this way we have identified $\widehat{X}_{[t]}$ with the product $X_{[t]}^p \times X_p$, where

$$X_{[t]}^p = \{x \in \mathrm{GL}_n^*(\mathbb{A}_f^p)/\widehat{\Gamma}_N^p \mid x^{-1}\gamma x \in \widehat{K}_N^p\},$$
$$X_p = \{x \in \mathrm{GL}_n^*(\mathbb{Q}_p)/\mathrm{GL}_n^*(\mathbb{Z}_p) \mid x^{-1}\alpha_q^{-1}\gamma x \in K_p\}.$$

In summary,

$$\begin{aligned} |S_\gamma(\mathbb{Q})\backslash X_{[t]}| &= \int_{S_\gamma(\mathbb{Q})\backslash \mathrm{GL}_n^*(\mathbb{A}_f)} \chi^p(x^{-1}\gamma x)\chi_p(x^{-1}\alpha_q^{-1}\gamma x)\, dx \\ &= \mathrm{vol}(S_\gamma(\mathbb{Q})\backslash S_\gamma(\mathbb{A}_f)) \cdot I_\gamma^p \cdot I_{\gamma,p}. \end{aligned}$$

□

9.9. If N is odd (with $N \geq 3$ and $p \nmid N$) then the formula must be modified slightly. The pairs $(\widehat{L}, \alpha)$ appear in $\mathrm{GL}_n^*(\mathbb{A}_f)$-orbits $X_{[t]}$ corresponding to cohomology classes $[t] \in H^1(\langle\tau_0\rangle, \widehat{K}_N^0)$ as before. However the class $[t]$ vanishes in $H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\mathbb{A}_f))$ (rather than in $H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\widehat{\mathbb{Z}}))$. Then $t = g^{-1}\tilde{g}$ for some $g \in \mathrm{Sp}_{2n}(\mathbb{A}_f)$ so the orbit $X_{[t]}$ is isomorphic to $\widehat{J}_{[t]} \backslash \mathrm{GL}_n^*(\mathbb{A}_f)$ where $\widehat{J}_{[t]} = \mathrm{GL}_n^*(\mathbb{A}_f) \cap g\widehat{K}_N g^{-1}$. Haar measure on $\mathrm{GL}_n^*(\mathbb{A}_f)$ should be chosen to give measure one to the set $\widehat{J}_{[t]}$, and the function χ^p in (9.8.1) should be replaced by the characteristic function ${}^{[t]}\chi^p$ on $\mathrm{GSp}_{2n}(\mathbb{A}_f^p)$ of the set $g\widehat{K}_N^p g^{-1}$.

9.10. ***Kottwitz integral.*** If we drop the involutions and real structures in the preceding sections then the same procedure as Proposition 9.4 identifies the number of isomorphism classes of Φ_ε-positive principally polarized Deligne modules with level N structure ($N \geq 3$ and $p \nmid N$) that are $\mathbb{Q}$-isogenous to (T_0, γ, ω_0) with the set $Z_\gamma(\mathbb{Q})\backslash Y$ where Y denotes the set of pairs $(\widehat{L}, \alpha)$ consisting of a lattice $\widehat{L} \subset \mathbb{A}_f^{2n}$, symplectic up to homothety and preserved by γ and by $q\gamma^{-1}$, and a level N structure α. As in [Kottwitz 1990], this gives

$$|Z_\gamma(\mathbb{Q})\backslash Y| = \mathrm{vol}(Z_\gamma(\mathbb{Q})\backslash Z_\gamma(\mathbb{A}_f)) \cdot \mathcal{O}_\gamma^p \cdot \mathcal{O}_{\gamma,p},$$

where

$$\mathcal{O}_\gamma^p = \int_{Z_\gamma(\mathbb{A}_f^p)\backslash G(\mathbb{A}_f^p)} f^p(g^{-1}\gamma g)\, dg \tag{9.10.1}$$

and

$$\mathcal{O}_{\gamma,p} = \int_{Z_\gamma(\mathbb{Q}_p)\backslash G(\mathbb{Q}_p)} f_p(g^{-1}\gamma g)\, dg = \int_{Z_\gamma(\mathbb{Q}_p)\backslash G(\mathbb{Q}_p)} \chi_p(g^{-1}\alpha_q^{-1}\gamma g)\, dg, \tag{9.10.2}$$

where f^p is the characteristic function on $G(\mathbb{A}_f^p)$ of $\widehat{K}_N^p$, and f_p is the characteristic function on $G(\mathbb{Q}_p)$ of $K_p\left(\begin{smallmatrix} I & 0 \\ 0 & qI \end{smallmatrix}\right)K_p$; see Lemma 9.6 and Section 9.10.

10. The counting formula

10.1. Fix a finite field $k = \mathbb{F}_q$ with q elements, and characteristic $p > 0$. Let $N \geq 3$ be a positive integer relatively prime to p. The theorem of Kottwitz specializes to:

Theorem 10.2 [Kottwitz 1990; 1992]. *The number $A(q)$ of principally polarized ordinary abelian varieties with principal level N structure, over the field $k = \mathbb{F}_q$, is finite and is equal to*

$$\sum_{\gamma_0} \sum_{\gamma \in \mathcal{C}(\gamma_0)} \mathrm{vol}(Z_\gamma(\mathbb{Q})\backslash Z_\gamma(\mathbb{A}_f)) \cdot \mathcal{O}_\gamma^p \cdot \mathcal{O}_{\gamma,p}, \tag{10.2.1}$$

where $\mathcal{O}_\gamma^p$ and $\mathcal{O}_{\gamma,p}$ are defined in (9.10.1) *and* (9.10.2).

10.3. ***Explanation and proof.*** Rather than start with the general formula of [Kottwitz 1990] and figure out what it says in the case of ordinary abelian varieties, we will follow the proof in [Kottwitz 1990], but apply it to Deligne modules; see also [Achter and Gordon 2017; Clozel 1993]. As discussed in the introduction, the result differs from the formula in [Kottwitz 1990] in two ways: (1) the invariant $\alpha(\gamma_0; \gamma, \delta)$ does not appear in our formula and (2) the twisted orbital integral in [Kottwitz 1990] (at p) is replaced by an ordinary orbital integral.

Recall Deligne's embedding $\varepsilon : W(\bar{k}) \to \mathbb{C}$. It determines a CM type Φ_ε on the CM algebra $\mathbb{Q}[F]$ for every Deligne module (T, F). As described in Section 3, Deligne constructs an equivalence of categories between the category of Φ_ε-positively polarized Deligne modules and the category of polarized ordinary abelian varieties over k, so we may count Deligne modules (that are Φ_ε-positively polarized) rather than abelian varieties.

The proof of (10.2.1) now follows five remarkable pages (pp. 203–207) in [Kottwitz 1990]. Roughly speaking, the first sum indexes the $\bar{\mathbb{Q}}$-isogeny classes, the second sum indexes the $\mathbb{Q}$-isogeny classes within a given $\bar{\mathbb{Q}}$-isogeny class, and the orbital integrals count the number of isomorphism classes within a given $\mathbb{Q}$-isogeny class.

10.4. The first sum is over rational representatives $\gamma_0 \in \mathrm{GSp}_{2n}(\mathbb{Q})$, one from each $\mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$-conjugacy class of semisimple elements such that the characteristic polynomial of γ_0 is an ordinary Weil q-polynomial (see Appendix A). Let $\mathcal{C} \subset \mathrm{GSp}_{2n}(\mathbb{Q})$ denote the $\mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$-conjugacy class of γ_0 within $\mathrm{GSp}_{2n}(\mathbb{Q})$. The first sum could equally well be considered as a sum over such conjugacy classes.

By Proposition 3.12, the set $\mathcal{C} \subset \mathrm{GSp}_{2n}(\mathbb{Q})$ contains elements that are *viable* with respect to the CM type Φ_ε (see Section 3.11), and the set of such Φ_ε-viable elements constitutes the intersection of $\mathcal{C}$ with a unique $\mathrm{Sp}_{2n}(\mathbb{R})$ conjugacy class. Therefore we may (and do) choose the representative $\gamma_0 \in \mathcal{C}$ to be Φ_ε-viable.

By Proposition 6.2, the choice of conjugacy class $\mathcal{C}$ corresponds to a $\bar{\mathbb{Q}}$-isogeny class of polarized Deligne modules. In fact, according to Lemma 6.3, since $\gamma_0 \in \mathcal{C}$ is Φ_ε-viable, there exists a polarized Deligne module of the form $(L_1, \gamma_0, \omega_0)$ where $L_1 \subset \mathbb{Q}^{2n}$ is a lattice such that

(a) L_1 is preserved by γ_0 and by $q\gamma_0^{-1}$ and

(b) the standard symplectic form ω_0 takes integral values on L_1.

The next step is to decompose the set of Φ_ε-viable elements in $\mathcal{C}$ into $\mathrm{GSp}_{2n}(\mathbb{Q})$ conjugacy classes. Thus, the second sum is over representatives $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$, one from each $\mathrm{GSp}_{2n}(\mathbb{Q})$-conjugacy class of elements such that

(1) γ, γ_0 are $\mathrm{GSp}_{2n}(\bar{\mathbb{Q}})$-conjugate (i.e., $\gamma \in \mathcal{C}$), and

(2) γ, γ_0 are $\mathrm{Sp}_{2n}(\mathbb{R})$-conjugate (i.e., γ is Φ_ε-viable).

Fix such an element γ. As explained in Section 8.3 (and Proposition 8.2) this choice of γ for the second sum corresponds to the choice of a $\mathbb{Q}$-isogeny class of Φ_ε-positively polarized Deligne modules within the $\overline{\mathbb{Q}}$-isogeny class of $(L_1, \gamma_0, \omega_0)$. The chosen element γ arises from some (not necessarily principally) Φ_ε-positively polarized Deligne module, say, (T_0, γ, ω_0) where $T_0 \subset \mathbb{Q}^{2n}$ is a lattice that also satisfies (a) and (b) above.

The set of isomorphism classes of Φ_ε-positive principally polarized Deligne modules within the $\mathbb{Q}$-isogeny class of (T_0, γ, ω_0) is identified, using Proposition 9.4 (see Section 9.10), with the quotient $Z_\gamma(\mathbb{Q})\backslash Y$, where $Z_\gamma(\mathbb{Q})$ is the centralizer of γ in $\mathrm{GSp}_{2n}(\mathbb{Q})$ and where Y denotes the set of pairs $(\widehat{L}, \alpha)$ consisting of a lattice $\widehat{L} \subset \mathbb{A}_f^{2n}$ and a level N structure α, satisfying (9.3.1) and (9.3.2), that is, $\widehat{L}$ is a lattice that is *symplectic up to homothety* (see Section D.4) and is preserved by γ and by $q\gamma^{-1}$, and the level structure is compatible with γ and with the symplectic structure. Decomposing the lattice $\widehat{L}$ into its adèlic components gives a product decomposition $Y \cong Y^p \times Y_p$ as described in Proposition 9.8 and Section 9.10. This in turn leads to the product of orbital integrals in (10.2.1).

Although the second sum in (10.2.1) may have infinitely many terms, only finitely many of the orbital integrals are nonzero. This is a consequence of Theorem 7.1, or of the more general result in [Kottwitz 1986, Proposition 8.2]. This completes the proof of Theorem 10.2. □

10.5. ***Counting real structures.*** Let τ_0 be the standard involution on $\mathbb{Q}^n \oplus \mathbb{Q}^n$ (see Appendix B). For $g \in \mathrm{GSp}_{2n}$, let $\tilde{g} = \tau_0 g \tau_0^{-1}$. Define $H = \mathrm{GL}_n^* \cong \mathrm{GL}_1 \times \mathrm{GL}_n$ to be the fixed point subgroup of this action, as in Section 5.4. If $\gamma \in \mathrm{GSp}_{2n}$, denote its H-centralizer by

$$S_\gamma = \{x \in \mathrm{GL}_n^* \mid x\gamma = \gamma x\}.$$

Assume the level $N \geq 3$ is even (see Section 9.9) and not divisible by p. Let χ^p denote the characteristic function of $\widehat{K}_N^p$ and let χ_p denote the characteristic function of $K_p = \mathrm{GSp}_{2n}(\mathbb{Z}_p)$.

Theorem 10.6. *The number of isomorphism classes of principally polarized ordinary abelian varieties with real structure is finite and is equal to*

$$\sum_{A_0} \sum_{C} |\widehat{H}^1| \operatorname{vol}(S_\gamma(\mathbb{Q})\backslash S_\gamma(\mathbb{A}_f)) \times \int_{S_\gamma(\mathbb{A}_f)\backslash H(\mathbb{A}_f)} \chi^p(x^{-1}\gamma x)\chi_p(x^{-1}\alpha_q^{-1}\gamma x)\, dx. \tag{10.6.1}$$

10.7. ***Explanation and proof.*** As in Theorem 10.2, the first sum indexes the $\overline{\mathbb{Q}}$-isogeny classes, the second sum indexes $\mathbb{Q}$-isogeny classes within a given $\overline{\mathbb{Q}}$-isogeny class, and the orbital integrals count the number of isomorphism classes within a $\mathbb{Q}$-isogeny class.

The first sum is over representatives, one from each $\mathrm{GL}_n(\mathbb{Q})$-conjugacy class (which is the same as the $\mathrm{GL}_n(\overline{\mathbb{Q}})$ conjugacy class) of semisimple elements $A_0 \in \mathrm{GL}_n(\mathbb{Q})$ whose characteristic polynomial $h(x) = b_0 + b_1 x + \cdots + x^n \in \mathbb{Z}[x]$ satisfies (see Appendix A)

(h1) $b_0 \neq 0$ and $p \nmid b_0$,

(h2) the roots $\beta_1, \beta_2, \ldots, \beta_n$ of h are totally real and $|\beta_i| < \sqrt{q}$ for $1 \leq i \leq n$.

By Proposition 6.2 the terms in this sum correspond to $\overline{\mathbb{Q}}$-isogeny classes of Φ_ε-positively polarized Deligne modules with real structure.

Fix such an element $A_0 \in \mathrm{GL}_n(\mathbb{Q})$. By Proposition 5.8, there exist C_0 such that the element

$$\gamma_0 = \begin{pmatrix} A_0 & B_0 \\ C_0 & {}^t A_0 \end{pmatrix} \in \mathrm{GSp}_{2n}(\mathbb{Q}),$$

(where $B_0 = (qI - A_0^2)C_0^{-1}$) is q-inversive (Section 5) and viable (Section 3.11) with respect to the CM type Φ_ε. (Viability corresponds to an appropriate choice of signature $\mathrm{sig}(A_0; C_0)$; see Proposition 5.8.) Then γ_0 corresponds to some Φ_ε-positively polarized Deligne module with real structure which (by Lemma 6.3) may be taken to be of the form $(L_1, \gamma_0, \omega_0, \tau_0)$ where $L_1 \subset \mathbb{Q}^{2n}$ is a lattice that is preserved by τ_0 and by γ_0 and $q\gamma_0^{-1}$.

The second sum in (10.6.1) is over representatives $C \in \mathrm{GL}_n(\mathbb{Q})$, one from each $Z_{\mathrm{GL}_n(\mathbb{Q})}(A_0)$-congruence class (Section 5.4) of matrices such that

(1) C is symmetric and nonsingular,

(2) $A_0 C = C\, {}^t A_0$,

(3) $\mathrm{sig}(A_0; C) = \mathrm{sig}(A_0; C_0)$ (see Section 5.4).

According to Proposition 8.2, the elements in this sum correspond to $\mathbb{Q}$-isogeny classes of Φ_ε-positively polarized Deligne modules with real structure that are in the same $\overline{\mathbb{Q}}$-isogeny class as $(L_1, \gamma_0, \omega_0, \tau_0)$. Let us fix such an element C and let $\gamma = \left(\begin{smallmatrix} A_0 & B \\ C & {}^t A \end{smallmatrix}\right)$ be the corresponding element from Proposition 8.2 (where $B = (A_0^2 - qI)C^{-1}$). Then γ is q-inversive and viable and it corresponds to some Φ_ε-positively polarized Deligne module with real structure, say $(T_0, \gamma, \omega_0, \tau_0)$ which we will use as a "basepoint" in the $\mathbb{Q}$-isogeny class determined by A_0, B.

(In fact, the first two sums may be replaced by a single sum over $\mathrm{GL}_n(\mathbb{Q})$-conjugacy classes of semisimple elements $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ that are q-inversive, whose characteristic polynomial is an ordinary Weil q-polynomial, and that are Φ_ε-viable.)

According to Proposition 9.4, the isomorphism classes of Φ_ε-positive principally polarized Deligne modules with real structure and level N structure that are $\mathbb{Q}$-isogenous to $(T_0, \gamma, \omega_0, \tau_0)$ correspond to isomorphism classes of pairs $(\widehat{L}, \alpha)$

(consisting of a lattice $\widehat{L} \subset \mathbb{A}_f^{2n}$ and a level structure) that satisfy (9.3.0), (9.3.1) and (9.3.2). In Proposition D.7 these lattices are divided into cohomology classes $[t] \in \widehat{H}^1 = H^1(\langle\tau_0\rangle, \widehat{K}_N^0)$. Each cohomology class provides the same contribution, which accounts for the factor of $|\widehat{H}^1|$. The number of isomorphism classes of pairs $(\widehat{L}, \alpha)$ corresponding to each cohomology class is proven, in Proposition 9.8, to equal the value of the orbital integral in (10.6.1).

The second sum in (10.6.1) (that is, the sum over C) may have infinitely many terms. However it follows from Theorem 7.1 that only finitely many of those terms give a nonzero contribution to the sum. This completes the proof of (10.6.1).

11. Totally real lattice modules

11.1. Suppose (T, F, τ) is a Deligne module of rank $2n$ over $\mathbb{F}_q$ with a real structure. The fixed point set or "real sublattice" $L = T^\tau$ has an interesting endomorphism[10] $A = (F + V) \mid L$, in which case the characteristic polynomial of A is $h(2x)$ where $h(x)$ is the real counterpart to the characteristic polynomial of F; see Section 5.1. Although it is not required for the rest of this paper, it is interesting to examine these structures in more detail.

If $\alpha : T/NT \to (\mathbb{Z}/N\mathbb{Z})^{2n}$ is a level N structure that is compatible with τ then its restriction to the fixed point set $\beta : L/NL \to (\mathbb{Z}/N\mathbb{Z})^n$ is a level N structure on (L, A). Thus, the category of Deligne modules (resp. with level N structure) fibers over a "totally real" category of lattices and endomorphisms (resp. with level N structure):

Definition 11.2. A totally real lattice module (of rank n and norm q) is a pair (L, A) where L is a free abelian group of rank n and $A : L \to L$ is a semisimple endomorphism whose eigenvalues α are totally real algebraic integers with $|\rho(\alpha)| \le \sqrt{q}$ for every embedding $\rho : \mathbb{Q}[\alpha] \to \mathbb{R}$. The module (L, A) is *ordinary* if $|\rho(\alpha)| < \sqrt{q}$ (for all eigenvalues α and all embeddings ρ) and $\det(A)$ is not divisible by p (see Proposition A.3). A level N structure on (L, A) is an isomorphism

$$\beta : L/NL \to (\mathbb{Z}/N\mathbb{Z})^n$$

such that $\beta \circ \bar{A} = \beta$ where $\bar{A} = A \pmod N$. A polarization (resp. principal polarization) of (L, A) is a symmetric bilinear form $R : L \times L \to \mathbb{Z}$ that is nonsingular over $\mathbb{Q}$ (resp. over $\mathbb{Z}$) such that $\mathbb{Q}[A]$ acts as an algebra of self adjoint operators on $L \otimes \mathbb{Q}$.

If $h(x)$ is the characteristic polynomial of an ordinary totally real lattice module then $h(2x)$ is an ordinary real Weil q-polynomial (see Appendix A).

[10] In this section, our use of the letter A differs by a factor of 2 from our previous use in Section 5.1 in the matrix representation for F as $\gamma = \left(\begin{smallmatrix} A & B \\ C & {}^tA \end{smallmatrix}\right) \in \mathrm{GSp}_{2n}(\mathbb{Q})$ to guarantee that the action of A preserves the lattice $L = T^\tau$.

Proposition 11.3. *The association* $(T, F, \tau) \mapsto (L = T^\tau, A = F + V)$ *defines a functor from the category of Deligne modules with real structure to the category of ordinary totally real lattice modules. It becomes an equivalence on the corresponding categories up to* $\mathbb{Q}$*-isogeny.*

Proof. In both cases the $\mathbb{Q}$-isogeny class is determined by the characteristic polynomial of A (see Proposition 6.2 below), so the result follows from Proposition A.3. □

(Similarly, if (T, F, ω, τ) is a Φ_ε-positively polarized Deligne module with real structure then a choice of totally positive imaginary algebraic integer $\iota \in \mathbb{Q}[F]$ determines a positive definite symmetric bilinear form $R(x, y) = \omega(x, \iota y)$ that takes integer values on $L = T^\tau$ such that $\mathbb{Q}[A]$ acts as an algebra of self adjoint operators on $L \otimes \mathbb{Q}$, thereby determining a polarized totally real lattice module $(L = T^\tau, A = F + V, R(x, y))$. However this procedure does not give an equivalence between the category of polarized Deligne modules and the category of polarized totally real lattice modules.)

11.4. Let $q = p^r$ and fix $n \geq 1$. Let $A \in \mathrm{GL}_n(\mathbb{Q})$ be a semisimple endomorphism whose determinant $\det(A)$ is not divisible by p, whose characteristic polynomial is integral, with roots α that are totally real such that $|\rho(\alpha)| < \sqrt{q}$ for every embedding $\rho : \mathbb{Q}[\alpha] \to \mathbb{R}$. Fix $N \geq 3$ not divisible by p. Let f be the characteristic function of the principal congruence subgroup of level N in $\mathrm{GL}_n(\mathbb{A}_f)$. Using arguments that are similar (but simpler) than those in Section 10, we find that *the number of isomorphism classes of* (*ordinary*) *totally real lattice modules of rank n, norm q and level N within the* $\mathbb{Q}$*-isogeny class determined by A is equal to the orbital integral*

$$\int_{Z_A(\mathbb{A}_f)\backslash GL_n(\mathbb{A}_f)} f(x^{-1}Ax)\, dx.$$

12. Further questions

12.1. We do not know whether the count of the number of "real" polarized Deligne modules has a rational zeta-function interpretation.

12.2. We do not know of a scheme-theoretic interpretation of antiholomorphic involution that applies to abelian varieties, rather than to Deligne modules. Consequently we do not know whether the notion of an antiholomorphic involution makes sense for general abelian varieties over $\mathbb{F}_q$. It would even be interesting to understand the case of supersingular elliptic curves.

12.3. In [Goresky and Tai 2003a], we showed that certain arithmetic hyperbolic 3-manifolds (and more generally, certain arithmetic quotients of quaternionic Siegel space) can be viewed as parametrizing abelian varieties with antiholomorphic multiplication by the integers $\mathcal{O}_d$ in a quadratic imaginary number field. It should

be possible to mimic these constructions using Deligne modules. Define an antiholomorphic multiplication on a Deligne module (T, F) by an order $\mathcal{O}$ in a CM field E to be a homomorphism $\mathcal{O} \to \mathrm{End}(T)$ such that each purely imaginary element $u \in \mathcal{O}$ acts in an antiholomorphic manner, that is, $uF = Vu$. One could probably count the number of isomorphism classes of principally polarized Deligne modules with level structure and with antiholomorphic multiplication by $\mathcal{O}$.

Appendix A: Weil polynomials and a real counterpart

A.1. Let π be an algebraic integer. It is *totally real* if $\rho(\pi) \in \mathbb{R}$ for every embedding $\rho : \mathbb{Q}(\pi) \to \mathbb{C}$. It is a *Weil q-integer* if $|\rho(\pi)|^2 = q$ for every embedding $\rho : \mathbb{Q}(\pi) \to \mathbb{C}$. (In this case, the field $\mathbb{Q}(\pi)$ is either a CM field, which is the usual case, or it is $\mathbb{Q}(\sqrt{q})$, the latter case occurring if and only if $\pi = \pm\sqrt{q}$.) A *Weil q-polynomial* is a monic polynomial $p(x) \in \mathbb{Z}[x]$ of even degree, all of whose roots are Weil q-integers. Let us say that a Weil q-polynomial $p(x) = \sum_{i=0}^{2n} a_i x^i$ is *ordinary* if the middle coefficient a_n is nonzero and is coprime to q. This implies that half of its roots in $\bar{\mathbb{Q}}_p$ are p-adic units and half of its roots are divisible by p; also that $x^2 \pm q$ is not a factor of $p(x)$, hence $p(x)$ has no roots in the set $\{\pm\sqrt{q}, \pm\sqrt{-q}\}$.

The characteristic polynomial of Frobenius associated to an abelian variety B of dimension n defined over the field $\mathbb{F}_q$ is a Weil q-polynomial. It is ordinary if and only if the variety B is ordinary; see Section 3.

A monic polynomial $p(x) \in \mathbb{Z}[x]$ is *totally real* if all of its roots are totally real algebraic integers. A *real* (resp. *real ordinary*) Weil q-polynomial of degree n is a monic polynomial $h(x) \in \mathbb{Z}[x]$ such that the polynomial $p(x) = x^n h(x + q/x)$ is a Weil q-polynomial (resp. an ordinary Weil q-polynomial). (See also [Howe and Lauter 2003; Howe and Lauter 2012]).

A.2. ***Real counterpart.*** Let $q \in \mathbb{Q}$. Let us say that a monic polynomial

$$p(x) = x^{2n} + a_{2n-1}x^{2n-1} + \cdots + a_0 \in \mathbb{C}[x]$$

is *q-palindromic* if it has even degree and if $a_{n-r} = q^r a_{n+r}$ for $1 \le r \le n$, or, equivalently, if

$$q^{-n}x^{2n}p\left(\frac{q}{x}\right) = p(x).$$

Thus $p(x)$ is q-palindromic if and only if the following holds: for every root π of $p(x)$ the number $q\pi^{-1}$ is also a root of $p(x)$. It is easy to see that every Weil q-polynomial is q-palindromic but the converse is not generally true. Let

$$p(x) = \prod_{j=1}^{n}(x - \alpha_j)\left(x - \frac{q}{\alpha_j}\right) = \sum_{i=0}^{2n} a_i x^i$$

be a q-palindromic polynomial with no real roots. Define the *associated real counterpart*

$$h(x) = \prod_{j=1}^{n} \left(x - \left(\alpha_j + \frac{q}{\alpha_j} \right) \right) = \sum_{i=0}^{n} b_i x^i$$

or equivalently, $p(x) = x^n h(x + q/x)$.

Proposition A.3. *Fix $n, q \in \mathbb{Z}$ with $n > 0$ and $q > 0$.*

(1) *Let $p(x) = \sum_{i=0}^{2n} a_i x^i \in \mathbb{C}[x]$ be q-palindromic with no real roots and let $h(x) = \sum_{j=0}^{n} b_j x^j \in \mathbb{C}[x]$ be its real counterpart. Then $p(x) \in \mathbb{Z}[x]$ if and only if $h(x) \in \mathbb{Z}[x]$.*

(2) *A q-palindromic polynomial $p(x) \in \mathbb{Z}[x]$ of even degree is a Weil q-polynomial if and only if the corresponding polynomial $h(x)$ is totally real.*

(3) *A totally real polynomial $h(x) \in \mathbb{Z}[x]$ is the real counterpart to a Weil q-polynomial $p(x)$ with no real roots if and only if the roots $\beta_1, \beta_2, \ldots, \beta_n \in \mathbb{R}$ of $h(x)$ satisfy $|\beta_i| < 2\sqrt{q}$ for $i = 1, 2, \ldots, n$.*

(4) *A Weil q-polynomial $p(x) \in \mathbb{Z}[x]$ is ordinary if and only if the constant coefficient $h(0) = b_0$ of the real counterpart is nonzero and is coprime to q. In this case, $p(x)$ is irreducible over $\mathbb{Q}$ if and only if $h(x)$ is irreducible over $\mathbb{Q}$.*

Proof. It is clear that $h \in \mathbb{Z}[x]$ implies $p \in \mathbb{Z}[x]$. Let $p(x) = \sum_{k=0}^{2n} a_k x^k \in \mathbb{C}[x]$ be a q-palindromic polynomial with roots $\alpha_i,\ q/\alpha_i$ for $1 \le i \le n$. The real counterpart is $h(x) = \sum_{j=0}^{n} b_j x^j = \prod_{i=1}^{n} (x - \beta_i)$ where $\beta_i = \alpha_i + q/\alpha_i$, hence

$$p(x) = x^n h\left(x + \frac{q}{x}\right) = \sum_{j=0}^{n} b_j \sum_{t=0}^{j} \binom{j}{t} q^{j-t} x^{n-j+2t}.$$

Set $r = n - j + 2t$. Then $n - j \le r \le n + j$ and $r - (n - j)$ is even, hence

$$p(x) = \sum_{r=0}^{2n} a_r x^r = \sum_{j=0}^{n} \sum_{r=n-j}^{n+j} A_{rj} b_j x^r,$$

where

$$A_{rj} = \binom{j}{(r + j - n)/2} q^{\frac{1}{2}(n-r+j)}$$

provided that $r + j - n$ is even and that $n - j \le r \le n + j$, and $A_{rj} = 0$ otherwise. Then $A_{n+s,s} = 1$ for all $1 \le s \le n$, so the lower half $A_{n+*,*}$ of the matrix A is nonsingular with determinant equal to 1. Let B be the inverse of the lower half of A. It is an integral matrix and for all $1 \le k \le n$,

$$b_k = \sum_{s=0}^{n} B_{ks} a_{n+s} \in \mathbb{Z}$$

which proves the first part of the proposition.

To verify statement (2), let $p(x)$ be a Weil q-polynomial. If it has any real roots then they must be of the form $\alpha = \pm\sqrt{q}$ so $\alpha + q/\alpha = \pm 2\sqrt{q}$ which is real. Every pair $\{\alpha, q/\alpha\}$ of complex roots are necessarily complex conjugate hence $\beta = \alpha + q/\alpha$ is real. Since $h(x)$ has integer coefficients this implies that every Galois conjugate of β is also real, hence $h(x)$ is a totally real polynomial. Conversely, given $p(x)$, if the associated polynomial $h(x)$ is totally real then for each root $\beta = \alpha + q/\alpha$ of $h(x)$, the corresponding pair of roots $\{\alpha, q/\alpha\}$ are both real or else they are complex conjugate, and if they are real then they are both equal to $\pm\sqrt{q}$. This implies that $p(x)$ is a Weil q-polynomial.

For part (3) of the proposition, each root $\beta_i \in \mathbb{R}$ of $h(x)$ is a sum $\beta_i = \alpha_i + q/\alpha_i$ of complex conjugate roots of $p(x)$. Hence α_i and q/α_i are the two roots of the quadratic equation

$$x^2 - \beta_i x + q = 0$$

which has real solutions if and only if $\beta_i^2 - 4q \geq 0$. Thus, $p(x)$ has no real roots if and only if $|\beta_i| < 2\sqrt{q}$ for $i = 1, 2, \ldots, n$.

For part (4), the polynomial $p(x)$ is ordinary if and only if exactly one of each pair of roots $\alpha_i, q/\alpha_i$ is a p-adic unit, from which it follows that each $\beta_i = \alpha_i + q/\alpha_i$ is a p-adic unit, hence the product $b_0 = \prod_{i=1}^n \beta_i$ is a p-adic unit (and it is nonzero). Conversely, if b_0 is a p-adic unit then so is each β_i so at least one of the elements in each pair $\alpha_i, q/\alpha_i$ is a unit. But in [Howe 1995; Deligne 1969] it is shown that this implies that exactly one of each pair of roots is a p-adic unit, so $p(x)$ is ordinary. The irreducibility statement follows from the formula $p(x) = x^n h(x + q/x)$. □

Lemma A.4. *Let $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ with multiplier $q \in \mathbb{Q}$. Then the characteristic polynomial $p(x)$ of γ is q-palindromic.*

Using the Jordan decomposition $\gamma = \gamma_s \gamma_u$ into semisimple and unipotent factors, it suffices to consider the case that γ is semisimple, so it can be diagonalized over $\overline{\mathbb{Q}}$, $\gamma = \left(\begin{smallmatrix} D & 0 \\ 0 & D' \end{smallmatrix}\right)$ where D and D') are diagonal matrices with $DD' = qI$ and entries $d_i' = q/d_i$. So $p(x) = \prod_{i=1}^n (x^2 - 2\alpha_i x + q)$ (where $\alpha_i = \frac{1}{2}(d_i + q/d_i)$) is a product of q-palindromic polynomials. □

Proposition A.5. *Let $p(x) = \sum_{i=0}^{2n} a_i x^i \in \mathbb{Q}[x]$ be a q-palindromic polynomial of degree $2n$ with no roots in the set $\{\pm\sqrt{q}, \pm\sqrt{-q}\}$. Then there exists a q-inversive element $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$ with multiplier q, whose characteristic polynomial is $p(x)$. Moreover, γ may be chosen to be semisimple, in which case it is uniquely determined up to conjugacy in $\mathrm{GSp}_{2n}(\overline{\mathbb{Q}})$ by its characteristic polynomial $p(x)$.*

Proof. Let $\lambda_1, \ldots, \lambda_n, q/\lambda_1, \ldots, q/\lambda_n$ denote the roots of $p(x)$. By assumption, λ_j and q/λ_j are distinct and their sum is nonzero. Factor the polynomial

$$p(x) = \prod_{i=1}^n (x - \lambda_i)\left(x - \frac{q}{\lambda_i}\right) = \prod_{i=1}^n \left(x^2 - \left(\lambda_i + \frac{q}{\lambda_i}\right)x + q\right),$$

set $\alpha_i = \frac{1}{2}(\lambda_i + q/\lambda_i)$ and define

$$h(x) = \prod_{i=1}^{n}(x - \alpha_i) = -h_0 - h_1 x - \cdots - h_{n-1}x^{n-1} + x^n.$$

(For convenience in this section, the signs of the coefficients of $h(x)$ have been modified from that of the preceding section.) The desired element is $\gamma = \left(\begin{smallmatrix} A & B \\ C & {}^tA \end{smallmatrix}\right)$ where the matrices A, B, C are defined as follows. The matrix A is the companion matrix for the polynomial $h(x)$, that is,

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & h_0 \\ 1 & 0 & 0 & \cdots & 0 & h_1 \\ 0 & 1 & 0 & \cdots & 0 & h_2 \\ 0 & 0 & 1 & \cdots & 0 & h_3 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1 & h_{n-1} \end{pmatrix}.$$

It is nonsingular (but not necessarily semisimple unless the roots of $h(x)$ are distinct). Now define

$$B = \begin{pmatrix} & & & & h_0 & 0 \\ & & & h_0 & h_1 & 0 \\ & & h_0 & h_1 & h_2 & 0 \\ & & & \cdots & & \\ h_0 & h_1 & h_2 & \cdots & h_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Then B is symmetric and nonsingular, and one checks directly that $AB = B\,{}^tA$. Define $C = B^{-1}(A^2 - qI)$ so that $A^2 - BC = qI$. These conditions guarantee that $\gamma \in \mathrm{GSp}_{2n}(\mathbb{Q})$, its multiplier is q, and it is q-inversive. Since the characteristic polynomial of A is $h(x)$, Lemma 5.2 implies that the characteristic polynomial of γ is $p(x)$.

If the roots of $p(x)$ are distinct then this element γ is semisimple. However if $p(x)$ has repeated roots it is necessary to proceed as follows. Factor $h(x) = \prod_{j=1}^{r} h_j^{m_j}(x)$ into its irreducible factors over $\mathbb{Q}$. This corresponds to a factorization $p(x) = \prod_{j=1}^{r} p_j^{m_j}(x)$ into q-palindromic factors. Take $A = \mathrm{diag}(A_1^{\times m_1}, \ldots, A_r^{\times m_r})$ to be a block-diagonal matrix with m_j copies of the matrix A_j. Then B, C will also be block-diagonal matrices, and γ will be the corresponding product of q-inversive symplectic matrices γ_j. It suffices to show that each nonzero γ_j is semisimple. Since $h_j(x)$ is irreducible over $\mathbb{Q}$, its roots are distinct, and the roots of $p_j(x)$ are the solutions to $x^2 - 2\alpha x + q = 0$ where $h_j(\alpha) = 0$. If $\pm\sqrt{q}$ is not a root of $h_j(x)$ then the roots of $p_j(x)$ are distinct, hence γ_j is semisimple. If $\pm\sqrt{q}$ is a root of $h_j(x)$ then $p(x) = (x - \sqrt{q})^2$ or $p(x) = (x^2 - q)^2$ depending on whether or not

$\sqrt{q} \in \mathbb{Q}$. In the first case we may take $A_j = \sqrt{q}$ and $B_j = C_j = 0$ and in the second case we may take $A_j = \left(\begin{smallmatrix} 0 & 1 \\ q & 0 \end{smallmatrix}\right)$ and $B_j = C_j = 0$. □

B: The symplectic group and its involutions

B.1. Let R be a commutative ring (with 1) and let T be a free, finite-dimensional R module. Let us say that an alternating form $\omega : T \times T \to R$ is *strongly nondegenerate*, if the induced mapping $\omega^\sharp : T \to \mathrm{Hom}_R(T, R)$ is an isomorphism.[11] Denote by $\mathrm{GSp}(T, \omega)$ the set of $g \in \mathrm{GL}(T)$ such that $\omega(gx, gy) = \lambda\omega(x, y)$ for some $\lambda = \lambda(g) \in R^\times$. Then λ is a character of $\mathrm{GSp}(T, \omega)$ and we say that $g \in \mathrm{GSp}(T, \omega)$ has *multiplier* $\lambda(g)$. The *standard symplectic form* on $T = R^{2n}$ is

$$\omega_0(x, y) = {}^t x J y \quad \text{where} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{B.1.1}$$

If $\omega : T \times T \to R$ is a symplectic form then a *symplectic basis* of T is an isomorphism $\Phi : T \to R^{2n}$ which takes ω to the standard symplectic form ω_0.

By abuse of notation we will write

$$\mathrm{GSp}_{2n}(R) = \mathrm{GSp}(R^{2n}, \omega_0) = \mathrm{GSp}(R^{2n}, J)$$

for the group of automorphisms of R^{2n} that preserve the standard symplectic form. If γ is in $\mathrm{GSp}_{2n}(R)$ then so is ${}^t\gamma^{-1}$, hence ${}^t\gamma$ is also. In this case, expressing γ as a block matrix $\gamma = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, the symplectic condition ${}^t\gamma J \gamma = qJ$ is equivalent to either of the following:

$${}^tAC,\ {}^tBD \ \text{ are symmetric and } {}^tAD - {}^tCB = qI. \tag{B.1.2}$$

$$A\,{}^tB,\ C\,{}^tD \ \text{ are symmetric and } A\,{}^tD - B\,{}^tC = qI. \tag{B.1.3}$$

Lemma B.2. *Let R be a principal ideal domain and let $\omega : T \times T \to R$ be a strongly nondegenerate symplectic form. Then T admits a symplectic basis. If $L', L'' \subset T$ are Lagrangian submodules such that $T = L' \oplus L''$ then the basis may be chosen so that L' and L'' are spanned by basis elements.*

Proof. Since ω is strongly nondegenerate there exist $x_1, y_1 \in T$ so that $\omega(x_1, y_1) = 1$. Let T_1 denote the span of x_1, y_1. Then $T = T_1 \oplus T_1^\perp$ because $T_1 \cap (T_1)^\perp = 0$ and for $v \in T$ we have that $u = v - \omega(v, y_1)x_1 - \omega(v, x_1)y_1 \in (T_1)^\perp$. So, T_1 and $T_1^\perp$ are projective, hence free, and ω is strongly nondegenerate on T_1. If $\dim(T) = 2$ we are done, otherwise strong nondegeneracy implies that ω induces an isomorphism

$$T_1 \oplus T_1^\perp \cong \mathrm{Hom}(T, R) \cong \mathrm{Hom}(T_1, R) \oplus \mathrm{Hom}(T_1^\perp, R).$$

Then $\omega \mid T_1$ is also strongly nondegenerate so it has a symplectic basis by induction.

[11] If R is an integral domain then an alternating form $B : T \times T \to R$ is *weakly nondegenerate* if $\omega^\sharp \otimes K$ is an isomorphism, where K is the fraction field of R.

If $T = L' \oplus L''$ is a decomposition into Lagrangian submodules, then the symplectic form induces an isomorphism $L'' \cong \mathrm{Hom}_R(L', R)$. Therefore an arbitrary basis of L' together with the dual basis of L'' will constitute a symplectic basis for T. □

B.3. Let R be a commutative ring. The *standard involution* $\tau_0 : R^{2n} \to R^{2n}$ is $\tau_0 = \left(\begin{smallmatrix} -I_n & 0 \\ 0 & I_n \end{smallmatrix}\right)$. If $g \in \mathrm{GSp}_{2n}(R)$, let $\tilde{g} = \tau_0^{-1} g \tau_0$; see Section 5.4.

Proposition B.4. *Let R be a principal ideal domain containing 2^{-1}. Let $\tau \in \mathrm{GSp}_{2n}(R)$ be an involution ($\tau^2 = I$) with multiplier -1. Then τ is $\mathrm{Sp}_{2n}(R)$-conjugate to τ_0.*

Proof. Write $T = R^{2n}$. The (standard) symplectic form ω_0 induces an isomorphism

$$T \cong \mathrm{Hom}(T, R), \quad \text{say, } x \mapsto x^\sharp. \tag{B.4.1}$$

Let T_+, T_- be the ± 1 eigenspaces of τ. Since $2^{-1} \in R$, any $x \in T$ may be written

$$x = \frac{x - \tau(x)}{2} + \frac{x + \tau(x)}{2} \in T_- + T_+$$

so $T = T_- \oplus T_+$. Therefore T_-, T_+ are projective, and hence free. Apply this splitting to (B.4.1) to find

$$\Phi : T_- \oplus T_+ \to \mathrm{Hom}(T_-, R) \oplus \mathrm{Hom}(T_+, R). \tag{B.4.2}$$

Since $\omega_0(\tau x, \tau y) = -\omega_0(x, y)$ it follows that $\Phi(x, y) = (y^\sharp, x^\sharp)$, so $\dim(T_-) = \dim(T_+) = n$ and we obtain an isomorphism $T_+ \cong \mathrm{Hom}(T_-, R)$. With respect to a basis of T_1 and the corresponding dual basis of T_+ the matrix of τ is $\left(\begin{smallmatrix} -I & 0 \\ 0 & I \end{smallmatrix}\right)$. □

The proposition fails if the ring R does not contain $\frac{1}{2}$, in fact we have:

Lemma B.5. *Let R be a Euclidean domain and let ω_0 be the standard symplectic form on R^{2n}. Let $\tau \in \mathrm{GSp}_{2n}(R)$ be an involution with multiplier equal to -1. Then τ is $\mathrm{Sp}_{2n}(R)$-conjugate to an element*

$$\begin{pmatrix} I & S \\ 0 & -I \end{pmatrix},$$

where S is a symmetric matrix consisting of zeroes and ones which may be taken to be one of the following: if $\mathrm{rank}(S) = r$ is odd then $S = \left(\begin{smallmatrix} I_r & 0 \\ 0 & 0 \end{smallmatrix}\right) = I_r \oplus 0_{n-r}$; if r is even then either $S = I_r \oplus 0_{n-r}$ or $S = H \oplus H \cdots \oplus H \oplus 0_{n-r}$ where $H = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ appears $r/2$ times in the sum.

Proof. There exists a vector $v \in \mathbb{Z}^{2n}$ that is primitive and has $\tau(v) = v$. By a lemma of Siegel (see [Freitag 1983, Satz A5.4]), there exists $g \in \mathrm{Sp}_{2n}(R)$ such that $gv = e_1 = (1, 0, \ldots, 0)$.

It follows that τ is $\mathrm{Sp}_{2n}(R)$-conjugate to a matrix $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ where

$$A = \begin{pmatrix} 1 & * \\ 0 & A_1 \end{pmatrix}, \quad B = \begin{pmatrix} * & * \\ * & B_1 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ 0 & C_1 \end{pmatrix}, \quad D = \begin{pmatrix} -1 & 0 \\ * & D_1 \end{pmatrix}$$

and where

$$\begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \in \mathrm{GSp}_{2n-2}(R)$$

is an involution with multiplier equal to -1. By induction, the involution τ is therefore conjugate to such an element where $A_1 = I$, B_1 is symmetric, $C_1 = 0$ and $D_1 = -I$. The condition $\tau^2 = I$ then implies that $A = I$, $D = -I$, $C = 0$ and B is symmetric. Conjugating τ by any element $\left(\begin{smallmatrix} I & T \\ 0 & I \end{smallmatrix}\right) \in \mathrm{Sp}_{2n}(R)$ (where T is symmetric) we see that B can be modified by the addition of an even number to any symmetric pair (b_{ij}, b_{ji}) of its entries. Therefore, we may take B to consist of zeroes and ones.

The problem may then be reduced to describing the list of possible symmetric bilinear forms on a $\mathbb{Z}/(2)$ vector space V, which are described in [Lidl and Niederreiter 1997, §6.2]. □

C: Positivity and $\mathbb{R}$-isogeny

Lemma C.1. *Let (T, F) be a Deligne module and let Φ be a CM type on $\mathbb{Q}[F]$* (see Section 3.7). *Let ω be a Φ-positive polarization of (T, F). Suppose $\beta \in \mathrm{End}_{\mathbb{Q}}(T, F)$ is a self isogeny that is fixed under the Rosati involution, that is, $\beta' = \beta$ where $\omega(\beta' x, y) = \omega(x, \beta y)$ for all $x, y \in T \otimes \mathbb{Q}$. Then there exists $\alpha \in \mathrm{End}_{\mathbb{Q}}(T, F) \otimes \bar{\mathbb{Q}}$ such that $\beta = \alpha'\alpha$. If $\beta' = \beta$ and β is positive definite then the element α may be chosen to lie in $\mathrm{End}_{\mathbb{Q}}(T, F) \otimes \mathbb{R}$. If ω_1, ω_2 are two Φ-positive polarizations of the same Deligne module (T, F) then there exists an $\mathbb{R}$-isogeny $(T, F, \omega_1) \to (T, F, \omega_2)$ with multiplier equal to* 1.

Proof. (See also [Kottwitz 1990, p. 206].) As indicated in [Mumford 1970, p. 220], the algebra $\mathrm{End}_{\mathbb{Q}}(T, F) \otimes \mathbb{R}$ is isomorphic to a product of matrix algebras $M_{d\times d}(\mathbb{C})$ such that $\beta' = {}^t\bar{\beta}$. Then $\beta' = \beta$ implies that β is Hermitian so there exists a unitary matrix $U \in M_{d\times d}(\mathbb{C})$ with $\beta = {}^t\bar{U} D U$ where D is a diagonal matrix of real numbers. Choose a square root $\sqrt{D} \in M_{d\times d}(\bar{\mathbb{Q}})$ and set $\alpha = \sqrt{D}U \in \mathrm{End}_{\mathbb{Q}}(T, F) \otimes \bar{\mathbb{Q}}$. Then $\alpha'\alpha = {}^t\bar{U} D U = \beta$ as claimed. Moreover, if β is positive definite then the entries of D are positive real numbers so we may arrange that $\sqrt{D} \in M_{d\times d}(\mathbb{R})$, so $\alpha \in \mathrm{End}_{\mathbb{Q}}(T, F) \times \mathbb{R}$ as claimed.

For the last sentence in the lemma, let $\beta \in \mathrm{End}_{\mathbb{Q}}(T, F)$ be the unique endomorphism so that $\omega_2(x, y) = \omega_1(\beta x, y)$. Then β is fixed under the Rosati involution for the polarization ω_1 because

$$\omega_1(\beta' x, y) = \omega_1(x, \beta y) = -\omega_1(\beta y, x) = -\omega_2(y, x) = \omega_2(x, y) = \omega_1(\beta x, y).$$

Moreover, β is positive definite: if $x \in T \otimes \mathbb{R}$ is an eigenvector of β with eigenvalue t then

$$t R_1(x, x) = R_1(\beta x, x) = \omega_1(\beta x, \iota x) = \omega_2(x, \iota x) = R_2(x, x) > 0$$

in the notation of Section 3.7. According to the first part of this lemma, there exists $\alpha \in \mathrm{End}_{\mathbb{Q}}(T, F) \otimes \mathbb{R}$ such that $\beta = \alpha'\alpha$, or

$$\omega_2(x, y) = \omega_1(\alpha'\alpha x, y) = \omega_1(\alpha x, \alpha y)$$

which says that α is an $\mathbb{R}$-isogeny which takes ω_1 to ω_2 with multiplier equal to 1. □

D: Symplectic cohomology

D.1. ***Nonabelian cohomology.*** Let R be a commutative ring with 1. As defined in Appendix B, the involution τ_0 of $R^n \times R^n$ is $\tau_0(x, y) = (-x, y)$. Let $\langle \tau_0 \rangle = \{1, \tau_0\} \cong \mathbb{Z}/(2)$ denote the group generated by the involution τ_0. For $g \in \mathrm{Sp}(2n, R)$ let $\tilde{g} = \tau_0 g \tau_0^{-1}$. This defines an action of the group $\langle \tau_0 \rangle$ on $\mathrm{Sp}(2n, R)$. Let $\Gamma \subset \mathrm{Sp}_{2n}(R)$ be a subgroup that is preserved by this action (that is, $\widetilde{\Gamma} = \Gamma$). Recall that a 1-cocycle for this action is a mapping $f : \langle \tau_0 \rangle \to \Gamma$ such that $f(1) = I$ and $f(\tau_0) = g$ where $g\tilde{g} = I$. We may write $f = f_g$ since the mapping f is determined by the element g. Then two cocycles f_g, $f_{g'}$ are cohomologous if there exists $h \in \Gamma$ such that $g' = h^{-1} g \tilde{h}$ or equivalently, such that $g' = \tilde{h} g h^{-1}$. The set of cohomology classes is denoted

$$H^1(\langle \tau_0 \rangle, \Gamma).$$

If $\tau \in \mathrm{GSp}_{2n}(R)$ is another involution (meaning that $\tau^2 = I$) with multiplier equal to -1 then $g = \tau\tau_0$ defines a cocycle since $g\tilde{g} = 1$. One easily checks the following.

Proposition D.2. *Let $\Gamma \subseteq \mathrm{Sp}_{2n}(R)$ be a subgroup that is normalized by τ_0. The mapping $\tau \mapsto \tau\tau_0$ determines a one-to-one correspondence between the set of Γ-conjugacy classes of involutions (i.e., elements of order 2), $\tau \in \Gamma.\tau_0$ and the cohomology set $H^1(\langle \tau_0 \rangle, \Gamma)$.*

D.3. ***Lattices and level structures.*** If $L \subset \mathbb{Q}^{2n}$ is a lattice, its *symplectic dual* is the lattice

$$L^\vee = \{x \in \mathbb{Q}^{2n} \mid \omega_0(x, y) \in \mathbb{Z} \text{ for all } y \in L\},$$

where ω_0 is the standard symplectic form. A lattice $L \subset \mathbb{Q}^{2n}$ is *symplectic* if $L^\vee = L$. A lattice $L \subset \mathbb{Q}^{2n}$ is *symplectic up to homothety* if there exists $c \in \mathbb{Q}^\times$ so that $L^\vee = cL$. In this case the symplectic form $b = c\omega_0$ is integer-valued and strongly nondegenerate on L. A lattice $L \subset \mathbb{Q}^{2n}$ is *real* if it is preserved by the standard involution τ_0, in which case write $\tau_L = \tau_0 \mid L$.

Fix $N \geq 1$ and let $\bar{L} = L/NL$. A level N structure on a lattice L is an isomorphism $\alpha : \bar{L} \to (\mathbb{Z}/N\mathbb{Z})^{2n}$. A level N structure α is compatible with an integer-valued symplectic form $b : L \times L \to \mathbb{Z}$ if $\alpha_*(b) = \bar{\omega}_0$ is the reduction modulo N of the standard symplectic form ω_0. A level N structure $\alpha : \bar{L} \to \bar{L}_0$ is *real* if it is compatible with the standard involution, that is, if $\bar{\tau}_0\alpha = \alpha\bar{\tau}_L : \bar{L} \to \bar{L}_0$.

D.4. ***Adèlic lattices.*** Let $\mathbb{A}_f = \prod'_{v<\infty} \mathbb{Q}_v$ (the restricted direct product) denote the finite adèles of $\mathbb{Q}$ and let $\widehat{\mathbb{Z}} = \prod_p \mathbb{Z}_p$. A $\widehat{Z}$-lattice $\widehat{M} \subset \mathbb{A}_f^{2n}$ is a product $\widehat{M} = \prod_{v<\infty} M_v$ of $\mathbb{Z}_v$-lattices $M_v \subset \mathbb{Q}_v^{2n}$ with $M_v = (\mathbb{Z}_v)^{2n}$ for almost all finite places v. If $\widehat{M} = \prod_{v<\infty} M_v$ is a lattice, its symplectic dual is $\widehat{M}^\vee = \prod_{v<\infty} M_v^\vee$ where

$$(M_v)^\vee = \{x \in \mathbb{Q}_v^{2n} \mid \omega_0(x, y) \in \mathbb{Z}_v \text{ for all } y \in M_v\}.$$

The lattice $\widehat{M}$ is *symplectic up to homothety* if there exists $\mathfrak{c} \in \mathbb{A}_f^\times$ such that $\widehat{M}^\vee = \mathfrak{c}\widehat{M}$. In this case, there exists $c \in \mathbb{Q}^\times$ (unique, up to multiplication by ± 1) such that $\widehat{M}^\vee = c\widehat{M}$, and the alternating form $b = c\omega_0$ takes $\widehat{Z}$ values on $\widehat{M}$. A lattice $\widehat{M}$ is *real* if it is preserved by the standard involution τ_0.

A level N structure on an adèlic lattice $\widehat{M}$ is an isomorphism

$$\beta : \widehat{M}/N\widehat{M} \to (\mathbb{Z}/N\mathbb{Z})^{2n}.$$

It is compatible with a $\widehat{Z}$-valued symplectic form $b : \widehat{M} \times \widehat{M} \to \widehat{Z}$ if $\beta_*(b) = \bar{\omega}_0$ is the reduction modulo N of the standard symplectic form. It is *real* if it commutes with the standard involution τ_0. The following statement is standard, see for example [Platonov and Rapinchuk 1994, Theorem 1.15]:

Lemma D.5. *Let $L \subset \mathbb{Q}^{2n}$ be a $\mathbb{Z}$-lattice and let $L_v = L \otimes \mathbb{Z}_v$ for each finite place v. Then*

- *$L_v = \mathbb{Z}_v^{2n}$ for almost all $v < \infty$.*
- *$L = \bigcap_{v<\infty}(\mathbb{Q}^{2n} \cap L_v)$.*
- *Given any collection of lattices $M_v \subset \mathbb{Q}_v^{2n}$ such that $M_v = \mathbb{Z}_v^{2n}$ for almost all $v < \infty$, there exists a unique $\mathbb{Z}$-lattice $M \subset \mathbb{Q}^{2n}$ such that $M_v = M \otimes \mathbb{Z}_v$ for all $v < \infty$.*

This correspondence is clearly compatible with symplectic structures, real structures and level structures.

D.6. ***The cohomology class of a symplectic lattice with "real" structure.*** Let $L \subset \mathbb{Q}^{2n}$ be a lattice, symplectic up to homothety (say, $L^\vee = cL$ where $c \in \mathbb{Q}$), and suppose that L is preserved by the standard involution $\tau_0 : \mathbb{Q}^{2n} \to \mathbb{Q}^{2n}$, in which case we refer to L as a "real" lattice. Let $\alpha : L/NL \to (\mathbb{Z}/N\mathbb{Z})^{2n}$ be a level N structure that is compatible with the involution (meaning that $\alpha_*(\bar{\tau}) = \bar{\tau}_0$ is the standard involution, where $\tau = \tau_0 \mid L$, and where the bar denotes reduction modulo N) and

with the nondegenerate symplectic form $b = c\omega_0$ on L (meaning that $\alpha_*(b) = \bar{\omega}_0$ is the standard symplectic form on $(\mathbb{Z}/N\mathbb{Z})^{2n}$). By the strong approximation theorem, the mapping

$$\mathrm{Sp}_{2n}(\mathbb{Z}) \to \mathrm{Sp}_{2n}(\mathbb{Z}/N\mathbb{Z})$$

is surjective. Together with the symplectic basis theorem (Lemma B.2) and the fact that α is compatible with $b = c\omega_0$, this implies that there exists $g \in \mathrm{GSp}_{2n}(\mathbb{Q})$ so that $(L, \alpha) = g.(L_0, \alpha_0)$, where $L_0 = \mathbb{Z}^{2n}$ is the standard lattice with its standard level N structure $\alpha_0 : L_0/NL_0 \to (\mathbb{Z}/N\mathbb{Z})^{2n}$. Both the lattice L and the level structure α are compatible with the involution which implies that $(L, \alpha) = g.(L_0, \alpha_0) = \tilde{g}.(L_0, \alpha_0)$ (where $\tilde{g} = \tau_0 g \tau_0^{-1}$). Therefore

$$t = g^{-1}\tilde{g} \in K_N^0 \subset \mathrm{Sp}_{2n}(\mathbb{Q})$$

is a cocycle (with multiplier equal to 1) which lies in the principal congruence subgroup

$$K_N^0 = \ker(\mathrm{Sp}_{2n}(\mathbb{Z}) \to \mathrm{Sp}_{2n}(\mathbb{Z}/N\mathbb{Z})).$$

Let $[(L, \alpha)] \in H^1(\langle\tau_0\rangle, K_N^0)$ denote the resulting cohomology class.

Similarly, an adèlic lattice $\widehat{L}$, symplectic up to homothety, and preserved by the involution τ_0, together with a level N structure β, (compatible with the involution and with the corresponding symplectic form) determine a cohomology class $[(\widehat{L}, \beta)] \in H^1(\langle\tau_0\rangle, \widehat{K}_N^0)$ where

$$\widehat{K}_N^0 = \ker(\mathrm{Sp}_{2n}(\widehat{\mathbb{Z}}) \to \mathrm{Sp}_{2n}(\mathbb{Z}/N\mathbb{Z})).$$

The following proposition is essentially the same as in [Rohlfs 1978].

Proposition D.7. *The resulting cohomology classes* $[(L, \alpha)]$ *and* $[(\widehat{L}, \beta)]$ *are well defined. The mapping* $L \mapsto \widehat{L} = \prod_v (L \otimes \mathbb{Z}_v))$ *determines a one-to-one correspondence between*

(1) $GL_n^*(\mathbb{Q})$*-orbits in the set of such pairs* (L, α) *that are symplectic up to homothety and compatible with the involution* (*as above*),

(2) $\mathrm{GL}_n^*(\mathbb{A}_f)$*-orbits in the set of such pairs* $(\widehat{L}, \beta)$ *that are symplectic up to homothety and compatible with the involution* (*as above*),

(3) *elements of the cohomology set*

$$H^1 := H^1(\langle\tau_0\rangle, K_N^0) \cong H^1(\langle\tau_0\rangle, \widehat{K}_N^0). \tag{D.7.1}$$

Proof. The cohomology class $[(L, \alpha)]$ is well defined: suppose that $(L, \alpha) = h.(L_0, \alpha_0)$ for some $h \in \mathrm{GSp}_{2n}(\mathbb{Q})$. Since L is symplectic up to homothety, the elements g, h have the same multiplier, hence $u = g^{-1}h \in K_N^0$. Therefore the cocycle $h^{-1}\tilde{h} = u^{-1}(g^{-1}\tilde{g})\tilde{u}$ is cohomologous to $g^{-1}\tilde{g}$.

Suppose $(L', \alpha') = g'.(L_0, \alpha_0)$ is another lattice with level N structure, with the same cohomology class. Then $(g')^{-1}\tilde{g}' = u^{-1}(g^{-1}\tilde{g})\tilde{u}$ for some $u \in K_N^0$ which implies that the element $h = g'u^{-1}g^{-1}$ is fixed under the involution. Hence $(L', \alpha') = h.(L, \alpha)$ is in the same $\mathrm{GL}_n^*(\mathbb{Q})$ orbit[12] as (L, α).

Similar remarks apply to adèlic lattices. Finally, Lemma D.5 implies that the cohomology sets (D.7.1) may be canonically identified. □

D.8. There is a simple relation between Propositions D.2 and D.7 which identifies the cohomology class of a lattice with a conjugacy class of involutions, as follows. Suppose (L, α) is a "real" symplectic (up to homothety) lattice with a level N structure. Express $(L, \alpha) = g.(L_0, \alpha_0)$ for some $g \in \mathrm{GSp}_{2n}(\mathbb{Q})$. Set $\tau = g^{-1}\tau_0 g = h^{-1}\tau_0 h$ where $h \in \mathrm{Sp}_{2n}(\mathbb{Q})$. Then τ is an involution in $K_N^0.\tau_0$ because $\tau\tau_0$ preserves (L_0, α_0), and the cohomology class of (L, α) coincides with the cohomology class of τ. We remark, moreover, if the cohomology class $[(L, \alpha)] \in H^1(\langle\tau_0\rangle, K_N^0)$ is trivial then the lattice L splits as a direct sum $L = L^+ \oplus L^-$ of ± 1 eigenspaces of τ and α determines a principal level N structure on each of the factors.

Proposition D.9. *If R is an integral domain containing $\frac{1}{2}$, then $H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(R))$ is trivial. If $2 \mid N$ the mapping $H^1(\langle\tau_0\rangle, K_N^0) \to H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\mathbb{Z}))$ is trivial. The cohomology sets*

$$H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\mathbb{Z})) \cong H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\widehat{\mathbb{Z}})) \cong H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\mathbb{Z}_2)) \tag{D.9.1}$$

are isomorphic and have order $(3n+1)/2$ if n is odd, or $(3n+2)/2$ if n is even.

Proof. By Proposition D.2, cohomology classes in $\mathrm{Sp}_{2n}(R)$ correspond to conjugacy classes of involutions with multiplier -1. If $\frac{1}{2} \in R$ then Proposition B.4 says there is a unique such involution, hence the cohomology is trivial. For the second statement, suppose $N \geq 2$ is even. Suppose $\alpha \in \mathrm{Sp}_{2n}(\langle\tau_0\rangle, K_N^0)$ is a cocycle. Then $\alpha\tau_0$ is an involution which, by Lemma B.5 implies that there exists $h \in \mathrm{Sp}_{2n}(\mathbb{Z})$ so that $h^{-1}\alpha\tilde{h} = \left(\begin{smallmatrix} I & B \\ 0 & I \end{smallmatrix}\right)$ where B is a symmetric matrix of zeroes and ones. It now suffices to show that $B = 0$ which follows from the fact that $\alpha \equiv I \pmod 2$ and that $h^{-1}\tilde{h} \equiv I \pmod 2$, so $B \equiv 0 \pmod 2$.

The cohomology set $H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\mathbb{Z}))$ is finite because it may be identified with $\mathrm{Sp}_{2n}(\mathbb{Z})$-conjugacy classes of involutions with multiplier -1 which, by Lemma B.5 corresponds to $\mathrm{GL}_n(\mathbb{Z})$-congruence classes of symmetric $n \times n$ matrices B consisting of zeroes and ones. Summing over the possible ranks $0 \leq r \leq n$ for the matrix B, with two possibilities when r is even and only one possibility when r is odd gives $(3n+1)/2$ for n odd and $(3n+2)/2$ for n even; see [Lidl and Niederreiter 1997]. The equation (D.9.1) holds since $\frac{1}{2} \in \mathbb{Z}_p$ for p odd. □

[12] So the orbit of (L, α) is isomorphic to $\Gamma_{(L,\alpha)} \backslash \mathrm{GL}_n^*(\mathbb{Q})$ where $\Gamma_{(L,\alpha)}$ is the stabilizer of (L, α).

Acknowledgments

We wish to thank R. Guralnick for useful conversations about the symplectic group and P. Deligne for his comments on the complex case. We would like to thank the editor, Don Blasius, and an anonymous referee for their suggestions and help with this paper. We are particularly grateful to the Defense Advanced Research Projects Agency for their support under grant no. HR0011-09-1-0010 and to our program officer at the time, Ben Mann. Goresky was also partially supported by the Institute for Advanced Study through a grant from the Simonyi Foundation. An earlier version of this paper (with more details) was posted at [Goresky and Tai 2017].

References

[Achter and Gordon 2017] J. D. Achter and J. Gordon, "Elliptic curves, random matrices and orbital integrals", *Pacific J. Math.* **286**:1 (2017), 1–24. MR Zbl

[Adler 1979] A. Adler, "Antiholomorphic involutions of analytic families of abelian varieties", *Trans. Amer. Math. Soc.* **254** (1979), 69–94. MR Zbl

[Andrianov 1987] A. N. Andrianov, *Quadratic forms and Hecke operators*, Grundlehren der Mathematischen Wissenschaften **286**, Springer, 1987. MR Zbl

[Barnet-Lamb et al. 2014] T. Barnet-Lamb, T. Gee, D. Geraghty, and R. Taylor, "Potential automorphy and change of weight", *Ann. of Math.* (2) **179**:2 (2014), 501–609. MR Zbl

[Borel 1969] A. Borel, *Introduction aux groupes arithmétiques*, Publications de l'Institut de Mathématique de l'Université de Strasbourg, XV. Actualités Scientifiques et Industrielles **1341**, Hermann, Paris, 1969. MR Zbl

[Clozel 1993] L. Clozel, "Nombre de points des variétés de Shimura sur un corps fini (d'après R. Kottwitz)", exposé 766, pp. 121–149 in *Séminaire Bourbaki,* 1992/93, Astérisque **216**, 1993. MR Zbl

[Comessatti 1926] A. Comessatti, "Sulle varietà abeliane reali", *Ann. Mat. Pura Appl.* **3**:1 (1926), 27–71. MR Zbl

[Deligne 1969] P. Deligne, "Variétés abéliennes ordinaires sur un corps fini", *Invent. Math.* **8** (1969), 238–243. MR Zbl

[Deligne 1973] P. Deligne, letter to Piatetski-Shapiro, 25 March 1973.

[Drinfeld 1976] V. G. Drinfeld, "Coverings of p-adic symmetric domains", *Funkcional. Anal. i Priložen.* **10**:2 (1976), 29–40. In Russian; translated in *Funct. Anal. Appl.* **10**:2 (1976), 107–115. MR Zbl

[Freitag 1983] E. Freitag, *Siegelsche Modulfunktionen*, Grundlehren der Mathematischen Wissenschaften **254**, Springer, 1983. MR Zbl

[Goresky and Tai 2003a] M. Goresky and Y. S. Tai, "Anti-holomorphic multiplication and a real algebraic modular variety", *J. Differential Geom.* **65**:3 (2003), 513–560. MR Zbl

[Goresky and Tai 2003b] M. Goresky and Y. S. Tai, "The moduli space of real abelian varieties with level structure", *Compositio Math.* **139**:1 (2003), 1–27. MR Zbl

[Goresky and Tai 2017] M. Goresky and Y.-s. Tai, "Real structures on ordinary Abelian varieties", preprint, 2017. arXiv

[Goresky and Tai 2019] M. Goresky and Y. S. Tai, "Real structures on polarized Dieudonné modules", *Pacific J. Math* **303**:1 (2019), 217–241.

[Gross and Harris 1981] B. H. Gross and J. Harris, "Real algebraic curves", *Ann. Sci. École Norm. Sup.* (4) **14**:2 (1981), 157–182. MR Zbl

[Harris and Taylor 2001] M. Harris and R. Taylor, *The geometry and cohomology of some simple Shimura varieties*, Annals of Mathematics Studies **151**, Princeton University Press, 2001. MR Zbl

[Harris et al. 2016] M. Harris, K.-W. Lan, R. Taylor, and J. Thorne, "On the rigid cohomology of certain Shimura varieties", *Res. Math. Sci.* **3** (2016), Paper No. 37, 308. MR Zbl

[Hooley 1958] C. Hooley, "On the representation of a number as the sum of a square and a product", *Math. Z.* **69** (1958), 211–227. MR Zbl

[Howe 1995] E. W. Howe, "Principally polarized ordinary abelian varieties over finite fields", *Trans. Amer. Math. Soc.* **347**:7 (1995), 2361–2401. MR Zbl

[Howe and Lauter 2003] E. W. Howe and K. E. Lauter, "Improved upper bounds for the number of points on curves over finite fields", *Ann. Inst. Fourier* (*Grenoble*) **53**:6 (2003), 1677–1737. MR Zbl

[Howe and Lauter 2012] E. W. Howe and K. E. Lauter, "New methods for bounding the number of points on curves over finite fields", pp. 173–212 in *Geometry and arithmetic*, edited by C. Faber et al., Eur. Math. Soc., Zürich, 2012. MR Zbl

[Humphreys 1975] J. E. Humphreys, *Linear algebraic groups*, Graduate Texts in Mathematics **21**, Springer, 1975. MR Zbl

[Ingham 1927] A. E. Ingham, "Some asymptotic formulae in the theory of numbers", *J. London Math. Soc.* **2**:3 (1927), 202–208. MR Zbl

[Katz 1981] N. Katz, "Serre–Tate local moduli", pp. 138–202 in *Algebraic surfaces* (Orsay, 1976–1978), edited by J. Giraud et al., Lecture Notes in Math. **868**, Springer, 1981. MR Zbl

[Knapp 2006] A. W. Knapp, *Basic algebra*, Birkhäuser, Boston, 2006. MR Zbl

[Kottwitz 1986] R. E. Kottwitz, "Stable trace formula: elliptic singular terms", *Math. Ann.* **275**:3 (1986), 365–399. MR Zbl

[Kottwitz 1990] R. E. Kottwitz, "Shimura varieties and λ-adic representations", pp. 161–209 in *Automorphic forms, Shimura varieties, and L-functions* (Ann Arbor, MI, 1988), vol. I, edited by L. Clozel and J. S. Milne, Perspect. Math. **10**, Academic Press, Boston, 1990. MR Zbl

[Kottwitz 1992] R. E. Kottwitz, "Points on some Shimura varieties over finite fields", *J. Amer. Math. Soc.* **5**:2 (1992), 373–444. MR Zbl

[Langlands and Rapoport 1987] R. P. Langlands and M. Rapoport, "Shimuravarietäten und Gerben", *J. Reine Angew. Math.* **378** (1987), 113–220. MR Zbl

[Lidl and Niederreiter 1997] R. Lidl and H. Niederreiter, *Finite fields*, vol. 20, Second ed., Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1997. MR Zbl

[Messing 1972] W. Messing, *The crystals associated to Barsotti–Tate groups: with applications to abelian schemes*, Lecture Notes in Mathematics **264**, Springer, 1972. MR Zbl

[Milne 2005] J. S. Milne, "Introduction to Shimura varieties", pp. 265–378 in *Harmonic analysis, the trace formula, and Shimura varieties*, edited by J. Arthur et al., Clay Math. Proc. **4**, American Mathematical Society, Providence, RI, 2005. MR Zbl

[Milne and Shih 1981] J. S. Milne and K.-y. Shih, "The action of complex conjugation on a Shimura variety", *Ann. of Math.* (2) **113**:3 (1981), 569–599. MR Zbl

[Mumford 1970] D. Mumford, *Abelian varieties*, Tata Institute of Fundamental Research Studies in Mathematics **5**, Oxford University Press, London, 1970. MR Zbl

[Nori and Srinivas 1987] M. V. Nori and V. Srinivas, "Canonical liftings", (1987). Appendix to V. B. Mehta and V. Srinivas, "Varieties in positive characteristic with trivial tangent bundle", *Compositio Math.* **64**:2 (1987), 191–212. MR Zbl

[Nygaard 1995] N. O. Nygaard, "Construction of some classes in the cohomology of Siegel modular threefolds", *Compositio Math.* **97**:1-2 (1995), 173–186. MR Zbl

[Patrikis and Taylor 2015] S. Patrikis and R. Taylor, "Automorphy and irreducibility of some l-adic representations", *Compos. Math.* **151**:2 (2015), 207–229. MR Zbl

[Platonov and Rapinchuk 1994] V. Platonov and A. Rapinchuk, *Algebraic groups and number theory*, Pure and Applied Mathematics **139**, Academic Press, Boston, 1994. MR Zbl

[Rohlfs 1978] J. Rohlfs, "Arithmetisch definierte Gruppen mit Galoisoperation", *Invent. Math.* **48**:2 (1978), 185–205. MR Zbl

[Scholze 2015] P. Scholze, "On torsion in the cohomology of locally symmetric varieties", *Ann. of Math.* (2) **182**:3 (2015), 945–1066. MR Zbl

[Seppälä and Silhol 1989] M. Seppälä and R. Silhol, "Moduli spaces for real algebraic curves and real abelian varieties", *Math. Z.* **201**:2 (1989), 151–165. MR Zbl

[Shimura 1975] G. Shimura, "On the real points of an arithmetic quotient of a bounded symmetric domain", *Math. Ann.* **215** (1975), 135–164. MR Zbl

[Shimura 1998] G. Shimura, *Abelian varieties with complex multiplication and modular functions*, Princeton Mathematical Series **46**, Princeton University Press, 1998. MR Zbl

[Shimura and Taniyama 1961] G. Shimura and Y. Taniyama, *Complex multiplication of abelian varieties and its applications to number theory*, Publications of the Mathematical Society of Japan **6**, Math. Soc. Japan, Tokyo, 1961. MR Zbl

[Silhol 1982] R. Silhol, "Real abelian varieties and the theory of Comessatti", *Math. Z.* **181**:3 (1982), 345–364. MR Zbl

[Spence 1972] E. Spence, "m-symplectic matrices", *Trans. Amer. Math. Soc.* **170** (1972), 447–457. MR Zbl

[Tate 1971] J. Tate, "Classes d'isogénie des variétés abéliennes sur un corps fini (d'après T. Honda)", exposé 352, pp. 95–110 in *Séminaire Bourbaki,* 1968/69, Lecture Notes in Math. **175**, Springer, 1971. MR Zbl

[Taylor 2008] R. Taylor, "Automorphy for some l-adic lifts of automorphic mod l Galois representations, II", *Publ. Math. Inst. Hautes Études Sci.* 108 (2008), 183–239. MR Zbl

Received August 22, 2018. Revised April 3, 2019.

MARK GORESKY
SCHOOL OF MATHEMATICS
INSTITUTE FOR ADVANCED STUDY
PRINCETON, NJ
UNITED STATES
goresky@ias.edu

YUNG SHENG TAI
DEPARTMENT OF MATHEMATICS
HAVERFORD COLLEGE
HAVERFORD, PA
UNITED STATES
ystai@comcast.net

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.217

REAL STRUCTURES ON POLARIZED DIEUDONNÉ MODULES

MARK GORESKY AND YUNG SHENG TAI

We define an "antiholomorphic involution" of a module M over the Dieudonné ring $\mathcal{E}(k)$ of a finite field k with $q = p^a$ elements to be an involution $\tau : M \to M$ that switches the action of $\mathcal{F}^a$ with that of $\mathcal{V}^a$. The definition extends to include quasi-polarizations of Dieudonné modules. Nontrivial examples exist. The number of isomorphism classes of quasi-polarized Dieudonné modules within a fixed isogeny class is shown to be given by a twisted orbital integral over the general linear group. Earlier (*Pacific J. Math.* 303:1 (2019), 165–215) we considered these notions in the case of ordinary abelian varieties over k, in which case the contribution at p to the number of isomorphism classes within an isogeny class was shown to be given by an ordinary orbital integral over the general linear group. The definitions here are shown to be equivalent to those in our previous paper and, as a consequence, the equality of the orbital integrals of both types is proven.

1. Introduction

Locally symmetric spaces associated to the group $\mathrm{GL}_n(\mathbb{R})$ for $n \geq 3$ do not carry a complex structure and do not admit an obvious reduction to characteristic $p > 0$. However, it is known ([Adler 1979; Gross and Harris 1981; Comessatti 1925; 1926; Goresky and Tai 2003a; 2003b; Milne and Shih 1981; Shimura 1975; Silhol 1982; Seppälä and Silhol 1989]) that such locally symmetric spaces parametrize *real* polarized abelian varieties (possibly with level structures). In an effort to find a characteristic p analog for such moduli spaces in [Goresky and Tai 2019] we introduced the notion of a *real structure* on an ordinary abelian variety A (or, rather, on its associated Deligne module T_A) defined over a finite field k: it is an "antiholomorphic" involution, that is, a linear involution that switches the action of the Frobenius and the Verschiebung. If A is the good, ordinary reduction of a CM variety $A/\mathbb{C}$ defined over $\mathbb{R}$ then complex conjugation of $A/\mathbb{C}$ induces such an involution on the Deligne module T_A. Over a finite field there are finitely many isomorphism classes of principally polarized ordinary abelian varieties with real structure and the number of isomorphism classes is given ([Goresky and Tai 2019])

MSC2010: 14G35, 16W10, 14K10, 22E27.
Keywords: Dieudonné module, abelian variety, real structure.

by a certain sum of orbital integrals over the general linear group $\mathrm{GL}_n \times \mathrm{GL}_1$. It is expected that these (or similar) definitions make sense beyond the "ordinary" case.

In Section 3.2, we extend the notion of a "real structure" to the case of (not necessarily ordinary) Dieudonné modules. We give examples (Section 3.3) to show that real structures often exist, even on supersingular Dieudonné modules. Then we show (Proposition 4.4) that the number of isomorphism classes of principally polarized "real" Dieudonné modules within a single isogeny class is given by a "twisted" orbital integral $TO(\delta)$ over the same general linear group $\mathrm{GL}_n \times \mathrm{GL}_1$.

We show that the constructions in this paper are compatible with those in [Goresky and Tai 2019], which requires an explicit description (Proposition 6.8) of the Dieudonné module (and its polarization) of an ordinary polarized abelian variety. Then we use this description to show (Proposition 6.12) that a real structure in the sense of [Goresky and Tai 2019] on an ordinary abelian variety determines a real structure (in the sense of this paper) on its Dieudonné module. This last step is not automatic: it requires a universal choice of involution on the Witt vectors, as constructed in Appendix A.

The compatibility between these two notions of real structure leads to a simplification of the twisted orbital integral $TO(\delta)$. The number of isomorphism classes of "real" Deligne modules (over $\mathbb{Z}_p$) is given by an (ordinary) orbital integral $O(\gamma)$: it is the component at p in the adèlic orbital integral of [Goresky and Tai 2019]. Using a linear algebra argument, we show (Section 7.5) that the orbital integral $O(\gamma)$ (which counts Deligne modules with real structure) coincides with the twisted orbital integral $TO(\delta)$ (which counts Dieudonné modules with real structure). This equality of orbital integrals is reminiscent of the results in [Kottwitz 1992] (for the symplectic group rather than the general linear group) in which the fundamental lemma for Levi subgroups is used in order to evaluate stable sums of twisted orbital integrals in terms of ordinary orbital integrals (and presumably a similar argument would work in our case as well).

2. Notation and terminology

Throughout this paper we fix a finite field $k = \mathbb{F}_q$ ($q = p^a$) of characteristic p. Let W denote the Witt ring functor, so that $W(k)$, $W(\bar{k})$ are the rings of (infinite) Witt vectors over $k, \bar{k}$, respectively, with fraction fields $K(k) = W(k) \otimes \mathbb{Q}_p$ and $K(\bar{k}) = W(\bar{k}) \otimes \mathbb{Q}_p$, respectively. We may identify $K(k)$ with the unique unramified extension of $\mathbb{Q}_p$ of degree $a = [k : \mathbb{F}_p]$. Let $W_0(\bar{k})$ denote the maximal unramified extension of $W(k)$. We may identify $W(\bar{k})$ with the completion of $W_0(\bar{k})$. Let $\sigma : W(\bar{k}) \to W(\bar{k})$ be the lift of the Frobenius mapping $\sigma : \bar{k} \to \bar{k}$, $\sigma(x) = x^p$ and let $\pi = \sigma^a$ be the topological generator for the Galois group $\mathrm{Gal}(\bar{k}/k) \cong \mathrm{Gal}(K(\bar{k})/K(k))$. Fix an identification, $\mathbb{Q}_p \cong K(\mathbb{F}_p)$ of the p-adic numbers with the fraction field of the Witt vectors of the prime field.

Let R be an integral domain with fraction field K. Let M be a free R-module of rank $2n$ and $V = M \otimes K$. An alternating bilinear form $\omega : M \times M \to R$ is *symplectic* if $\omega \otimes K : V \otimes V \to K$ is nondegenerate. It is *strongly nondegenerate* if the resulting $M \to \mathrm{Hom}_R(M, R)$ is an isomorphism. It is *symplectic up to homothety* if there exists $c \in K^\times$ such that $c\omega$ is strongly nondegenerate. The *standard symplectic form* ω_0 on $R^{2n} \times R^{2n}$ is that whose matrix is $J = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$. Set $G = \mathrm{GSp}_{2n}$ and for convenience denote

$$\Gamma_p = G(\mathbb{Z}_p) \quad \text{and} \quad \Gamma_W = G(W(k)). \tag{2.0.1}$$

The *standard involution* $\tau_0 : R^{2n} \to R^{2n}$ is the linear map with matrix $\left(\begin{smallmatrix} -I & 0 \\ 0 & I \end{smallmatrix}\right)$. Conjugation by τ_0, which we denote by

$$g \mapsto \tilde{g} = \tau_0 g \tau_0^{-1}$$

defines an action of the group $\langle \tau_0 \rangle \cong \mathbb{Z}/2\mathbb{Z}$ on GSp_{2n}. If $2 \in K^\times$ the fixed subgroup is

$$H = \mathrm{GL}_n^* = \left\{ \begin{pmatrix} A & 0 \\ 0 & \lambda\, {}^tA^{-1} \end{pmatrix} \in \mathrm{GSp}_{2n} \;\middle|\; A \in \mathrm{GL}_n;\, \lambda \in \mathbb{G}_m \right\} \cong \mathrm{GL}_n \times \mathbb{G}_m. \tag{2.0.2}$$

If $\mathcal{C}$ is a $\mathbb{Z}$-linear category then the associated category up to R isogeny is the category $\mathcal{C} \otimes R$ with the same objects but with morphisms $\mathrm{Hom}_{\mathcal{C}\otimes R}(x, y) = \mathrm{Hom}_{\mathcal{C}}(x, y) \otimes R$.

3. Dieudonné modules

3.1. ***Notation.*** Let $\mathcal{E} = \mathcal{E}(k)$ denote the Cartier–Dieudonné ring, that is, the ring of noncommutative $W(k)$-polynomials in two variables $\mathcal{F}$, $\mathcal{V}$, subject to the relations $\mathcal{F}(wx) = \sigma(w)\mathcal{F}(x)$, $\mathcal{V}(wx) = \sigma^{-1}(w)\mathcal{V}(x)$, and $\mathcal{F}\mathcal{V} = \mathcal{V}\mathcal{F} = p$, where $w \in W(k)$ and $x \in \mathcal{E}$. A *Dieudonné module* M is a module over the ring $\mathcal{E}(k)$ that is free and finite rank over $W(k)$.

The *covariant Dieudonné functor* (see, for example, [Chai et al. 2014, §B.3.5.6] or [Goren 2002, p. 245] or [Pink 2005]) assigns to each p-divisible group

$$G = \ldots \underset{\leftarrow}{\hookrightarrow} G_r \underset{\leftarrow}{\hookrightarrow} G_{r+1} \underset{\leftarrow}{\hookrightarrow} \ldots$$

a corresponding module $M(G) = \varprojlim M(G_r)$ over the Dieudonné ring $\mathcal{E}$.

A *quasi-polarization* (in the sense of [Moonen 2001; Oort 2001] and [Li and Oort 1998, §5.9] following [Oda 1969, p. 101]) of a Dieudonné module M is an alternating $W(k)$-bilinear form $\omega : M \times M \to W(k)$ such that $\omega \otimes K(k)$ is nondegenerate and $\omega(\mathcal{F}x, y) = \sigma\omega(x, \mathcal{V}y)$. (The use of "quasi" reflects the fact that there is no p-adic counterpart to the "positivity" condition found in the definition of a polarization for abelian varieties.) A $K(k)$-*isogeny* of polarized Dieudonné modules $(M, \omega) \to (M', \omega')$ is an element $\phi \in \mathrm{Hom}_{\mathcal{E}}(M, M') \otimes K(k)$ such that $\phi^*(\omega') = c\omega$ for some $c \in K(k)^\times$.

3.2. ***Real structures.*** Let M be a Dieudonné module of finite rank over $W(k)$ (with $k = \mathbb{F}_q$; $q = p^a$). Let ω be a quasi-polarization on M. Define a real structure on (M, ω) to be a $W(k)$-linear mapping $\tau_p : M \to M$ such that for all $x, y \in M$,

$$\tau_p^2 = I, \quad \tau_p \mathcal{F}^a \tau_p^{-1} = \mathcal{V}^a, \quad \omega(\tau_p x, \tau_p y) = -\omega(x, y). \tag{3.2.1}$$

As in [Kottwitz 1990, §12] the action of $\mathcal{F}$ may be expressed as $\delta\sigma$ for some $\delta \in \mathrm{GSp}(M \otimes K(k), \omega)$, so its norm

$$N(\delta) = \delta\sigma(\delta) \cdots \sigma^{a-1}(\delta) \in \mathrm{GSp}(M \otimes K(k), \omega)$$

coincides with the $W(k)$-linear action of $\mathcal{F}^a$. The second condition in (3.2.1) gives

$$\tau_p N(\delta) \tau_p^{-1} = q N(\delta)^{-1}.$$

3.3. ***Manin modules.*** Following [Manin 1963], let us define Dieudonné modules

$$M_{r,s} = \mathcal{E}(k)/\mathcal{E}(k)(\mathcal{F}^r + \mathcal{V}^s)$$

for nonnegative integers r, s. If $\bar{k}$ is an algebraic closure of k and if we extend scalars to

$$\overline{\mathcal{E}}(\bar{k}) = W(\bar{k})\Big[\frac{1}{p}\Big] \otimes \mathcal{E}(k),$$

it is shown in [Manin 1963] that if $\gcd(r, s) = 1$, the resulting modules $\overline{\mathcal{E}}(\bar{k}) \otimes_{\mathcal{E}(k)} M_{r,s}$ are simple and they account for all the simple Dieudonné modules. Elements of $M_{r,s}$ may be represented by (noncommutative) polynomials

$$x = \sum_{i=1}^{s-1} a_{-i} \mathcal{V}^i + a_0 + \sum_{j=1}^{r} a_j \mathcal{F}^j$$

(with $a_t \in W(k)$ and with identifications $\mathcal{F}^r = -\mathcal{V}^s$).

In the following paragraphs we will show that *the Manin modules $M_{r,s} \oplus M_{s,r}$ and the Manin modules $M_{r,r}$ admit quasi-polarizations and real structures.*

First suppose $r \neq s$. The elements $\{1, \mathcal{F}^j, \mathcal{V}^i\}$ $(1 \leq j \leq r;\ 1 \leq i \leq s-1)$ form a basis of $M_{r,s}$ over $W(k)$. The module $M_{s,r}$ admits a dual basis by setting

$$(\mathcal{F}^i)^\vee = \mathcal{V}^{r-i}, \quad (\mathcal{V}^j)^\vee = \mathcal{F}^{s-j}.$$

This gives rise to a $W(k)$-linear pairing $T : M_{r,s} \times M_{s,r} \to W(k)$ with

$$T(\mathcal{F}^i, \mathcal{V}^j) = \begin{cases} 1 & \text{if } i + j = r, \\ 0 & \text{otherwise,} \end{cases} \qquad T(\mathcal{V}^i, \mathcal{F}^j) = \begin{cases} 1 & \text{if } i + j = s, \\ 0 & \text{otherwise,} \end{cases}$$

such that $T(\mathcal{F}x, y) = \sigma(T(x, \mathcal{V}y))$. It follows that the alternating bilinear form

$$\omega(x \oplus y, x' \oplus y') = T(x, y') - T(x', y)$$

defines a quasi-polarization on $M_{r,s} \oplus M_{s,r}$. A real structure on this sum is defined by

switching the factors and exchanging $\mathcal{F}$ with $\mathcal{V}$. Explicitly, define $\tau_p : M_{r,s} \to M_{s,r}$ by

$$\tau_p\left(\sum_{i=1}^{s-1} a_{-i}\mathcal{V}^i + a_0 + \sum_{j=1}^{r} a_j\mathcal{F}^j\right) = \sum_{i=1}^{s-1} a_{-i}\mathcal{F}^i + a_0 + \sum_{j=1}^{r} a_j\mathcal{V}^j$$

and similarly for $\tau_p : M_{s,r} \to M_{r,s}$. Then $\tau_p^2 = I$ and

$$\tau_p(\mathcal{F}(x \oplus y)) = \sigma^2\mathcal{V}(\tau_p(x \oplus y))$$

which implies that $\tau_p\mathcal{F}^a = \mathcal{V}^a\tau_p$. Finally, one verifies for $x, y \in M_{r,s}$ and $x', y' \in M_{s,r}$ that

$$\omega(\tau_p(x \oplus y), \tau_p(x' \oplus y')) = -\omega(x \oplus y, x' \oplus y').$$

Now suppose $r = s$. The Manin module

$$M'_{r,r} = \mathcal{E}(k)/\mathcal{E}(k)(\mathcal{F}^r + \mathcal{V}^r)$$

is the Dieudonné module of a supersingular abelian variety. It has a $W(k)$-basis consisting of $\{\mathcal{V}^i,\ \mathcal{F}^j,\ \mathcal{V}^0 = \mathcal{F}^0 = 1,\ \mathcal{V}^r = -\mathcal{F}^r\}$ with $1 \le i, j \le r-1$. It admits a quasi-polarization which for $0 \le i, j \le r$ is well defined as

$$\omega(\mathcal{V}^i, \mathcal{F}^j) = \begin{cases} 1 & \text{if } i + j = r, \\ 0 & \text{otherwise,} \end{cases} \qquad \omega(\mathcal{F}^i, \mathcal{V}^j) = \begin{cases} -1 & \text{if } i + j = r, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\omega(x, y) = -\omega(y, x)$ and $\omega(\mathcal{F}x, y) = \sigma\omega(x, \mathcal{V}y)$ for all $x, y \in M'_{1,1}$. This module admits a real structure by setting $\tau_p(t\mathcal{F}^i) = t\mathcal{V}^i$ for $t \in W(k)$ and $0 \le i \le r$ (and in particular, $\tau_p(t\mathcal{F}^r) = -t\mathcal{F}^r$). It is easy to check that $\tau_p(\mathcal{F}^a x) = \mathcal{V}^a\tau_p(x)$ for all $a \ge 0$ and all $x \in M'_{r,r}$.

3.4. In [Manin 1963] the isogenous module $\mathcal{E}(k)/\mathcal{E}(k)(\mathcal{F}^r - \mathcal{V}^s)$ is used to replace the module $M_{r,s}$. However the "+" sign in the preceding example is crucial.

4. Counting Dieudonné modules

As in (2.0.1) let $\Gamma_W = G(W(k))$ with the standard symplectic form $\omega_0 = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$. Let $I_p = \left(\begin{smallmatrix} I & 0 \\ 0 & pI \end{smallmatrix}\right)$. By the theory of Smith normal form for the symplectic group (see [Spence 1972] or [Andrianov 1987, Lemma 3.3.6]), or by the Cartan decomposition for p-adic groups, we have the following:

Lemma 4.1. *Let $L_0 = W(k)^{2n} \subset K(k)^{2n}$ denote the standard lattice. Let $L \subset K(k)^{2n}$ be a $W(k)$-lattice. Then $L = hL_0$ for some $h \in G(K(k))$ and the following statements are equivalent.*

(1) $pL_0 \subset hL_0 \subset L_0$.

(2) $hL_0 \subset L_0,\ ph^{-1}L_0 \subset L_0$.

(3) $h \in \Gamma_W I_p \Gamma_W$. □

4.2. Assume $p \neq 2$. In this section we fix a Dieudonné module $(M, \mathcal{F}, \mathcal{V})$ with a quasi-polarization ω_M and a real structure $\tau_M : M \to M$. Then M is a free module over $W(k)$ of some even rank, say $2n$. Let $M_{\mathbb{Q}} = M \otimes K(k)$. We wish to understand the set X_M of (real) isomorphism classes of principally (quasi-)polarized Dieudonné modules that are $K(k)$-isogenous to M. In Proposition 4.4 below we show that the cardinality $|X_M|$ is given by a twisted orbital integral over the group $H \cong \mathrm{GL}_n \times \mathrm{GL}_1$ of (2.0.2).

Following the method of [Kottwitz 1990] let $\mathcal{X}_M$ denote the set of isomorphism classes in the category $\mathcal{C}_M$ whose objects consist of tuples $(P, \omega_P, \psi, \tau_P)$ where P is a Dieudonné module, ω_P is a principal quasi-polarization of P, where τ_P is a real structure on P and where $\psi \in \mathrm{Hom}_{W(k)}(P, M) \otimes K(k)$ is a $K(k)$ isogeny (meaning that $\psi \otimes K(k) : P_{\mathbb{Q}} \to M_{\mathbb{Q}}$ is an isomorphism) that commutes with $\mathcal{F}$, $\mathcal{V}$, takes τ_P to τ_M and satisfies $\psi^*(\omega_M) = c\omega_P$ for some $c \in K(k)^\times$. A morphism $\phi : P \to P'$ between left $\mathcal{E}(k)$ modules is in $\mathcal{C}_M$ if it is compatible with ω up to scalars, and it commutes with $\mathcal{F}, \mathcal{V}$ and the involutions $\tau_P . \tau_{P'}$. So there is a natural identification

$$X_M \cong I(M) \backslash \mathcal{X}_M,$$

where $I(M)$ denotes the group of $K(k)$ self-isogenies of (M, ω_M, τ_M).

4.3. The mapping $(P, \omega_P, \psi, \tau_P) \mapsto L = \psi(P)$ determines an identification between the set $\mathcal{X}_M$ and the set of $W(k)$-lattices $L \subset M_{\mathbb{Q}}$ that are preserved by $\mathcal{F}_M, \mathcal{V}_M, \tau_M$ and such that L is *symplectic up to homothety* meaning that $L^\vee = cL$ for some $c \in K(k)^\times$, where

$$L^\vee = \{x \in M_{\mathbb{Q}} \mid \omega_M(x, y) \in W(k) \text{ for all } y \in L\}.$$

By [Goresky and Tai 2019, Proposition B.4] there exists a $K(k)$-linear isomorphism $M_{\mathbb{Q}} \to K(k)^{2n}$ which takes the quasi-polarization ω_M to the standard symplectic form ω_0 and takes the involution τ_M to the standard involution $\tau_0 = \left(\begin{smallmatrix} -I & 0 \\ 0 & I \end{smallmatrix}\right) \in G(\mathbb{Z})$. From Section 3.2 the action of $\mathcal{F} \circ \sigma^{-1}$ becomes some element $\delta \in G(K(k))$ with multiplier p, that is well defined up to σ-conjugacy. The group $I(M)$ of self-isogenies becomes identified with the twisted centralizer (note that $\delta \notin H(K(k))$):

$$S_\delta(K(k)) = \{z \in H(K(k)) \mid z^{-1}\delta\sigma(z) = \delta\}.$$

Normalize the Haar measure on $H(K(k))$ so that $H(W(k))$ has volume one.

Proposition 4.4. *The choice of isomorphism $M_{\mathbb{Q}} \to K(k)^{2n}$ determines a one-to-one correspondence between the set of lattices $L \subset M_{\mathbb{Q}}$, symplectic up to homothety, that are preserved by $\mathcal{F}, \mathcal{V}, \tau_M$ and the set*

$$\{g \in H(K(k))/H(W(k)) \mid g^{-1}\delta\sigma(g) \in \Gamma_W I_p \Gamma_W\}. \tag{4.4.1}$$

Consequently the number of isomorphism classes

$$|X_M| = |S_\delta(K(k))\backslash \mathcal{X}_M|$$

of principally quasi-polarized real Dieudonné modules within the $K(k)$-isogeny class of M is given by the twisted orbital integral over $H = \mathrm{GL}_n^*$,

$$TO(\delta) = \int_{S_\delta(K)\backslash H(K)} \kappa_W(g^{-1}\delta\sigma(g))\, dg, \tag{4.4.2}$$

where κ_W is the characteristic function of $\Gamma_W I_p \Gamma_W \subset G(K(k))$.

Proof. Let $L_0 = W(k)^{2n} \subset K(k)^{2n}$ be the standard lattice. If $L \subset K(k)^{2n}$ is a $W(k)$-lattice, symplectic up to homothety, then $L = gL_0$ for some $g \in G(K(k))$. If it is preserved by $\mathcal{F}$, $\mathcal{V}$ then

$$pL_0 \subset g^{-1}\delta\sigma(g)L_0 \subset L_0 \tag{4.4.3}$$

which, by Lemma 4.1, is equivalent to $g^{-1}\delta\sigma(g) \in \Gamma_W I_p \Gamma_W$. (In the case of an "ordinary" Dieudonné module, a simpler formula holds; see Proposition 7.3).

If the lattice L is also preserved by the involution τ_0 then $g^{-1}g L_0 = L_0$ so that $\alpha = g^{-1}\tilde{g}$ is a 1-cocycle, defining a class in $H^1(\langle\tau_0\rangle, G(W(k)))$, which is trivial by [Goresky and Tai 2019, Proposition B.4] since $p \neq 2$. Thus, there exists $h \in G(W(k))$ so that $g^{-1}\tilde{g} = h^{-1}\tilde{h}$, hence $g' = gh^{-1} \in H(K(k)) = \mathrm{GL}_n^*(K(k))$ and $L = g'L_0$. Thus we may assume that $g \in H(K(k))$, while elements of $H(W(k))$ act trivially on the homothety class of the lattice L_0. If we normalize Haar measure so that $H(W(k))$ has volume one then the number of such lattices is given by the integral in (4.4.2). □

5. Deligne modules and ordinary abelian varieties

5.1. Recall from [Deligne 1969] that a *Deligne module* of rank $2n$ over the field $k = \mathbb{F}_q$ of q elements is a pair (T, F) where T is a free $\mathbb{Z}$-module of dimension $2n$ and $F : T \to T$ is an endomorphism such that the following conditions are satisfied:

(1) The mapping F is semisimple and all of its eigenvalues in $\mathbb{C}$ have magnitude $\sqrt{q}$.

(2) Exactly half of the eigenvalues of F in $\overline{\mathbb{Q}}_p$ are p-adic units and half of the eigenvalues are divisible by q.

(3) The middle coefficient of the characteristic polynomial of F is coprime to p.

(4) There exists an endomorphism $V : T \to T$ such that $FV = VF = q$.

A morphism $(T_A, F_A) \to (T_B, F_B)$ of Deligne modules is a group homomorphism $\phi : T_A \to T_B$ such that $F_B\phi = \phi F_A$. A polarization ([Howe 1995]) of a Deligne module (T, F) is a symplectic form $\omega : T \times T \to \mathbb{Z}$ (alternating and nondegenerate over $\mathbb{Q}$) such that $\omega(Fx, y) = \omega(x, Vy)$ for all $x, y \in T$, and such that the form

$R(x, y) = \omega(x, \iota y)$ is symmetric and positive definite, where ι is some (and hence, any) totally positive imaginary element of $\mathbb{Q}[F]$ (see [Howe 1995, §4.7]).

5.2. Following [Deligne 1969], for the rest of this paper we fix an embedding

$$\varepsilon : W(\bar{k}) \to \mathbb{C}. \tag{5.2.1}$$

By a theorem of Serre and Tate, [Drinfeld 1976; Katz 1981; Messing 1972; Nori and Srinivas 1987] an ordinary abelian variety A/k has a canonical lift $\bar{A}$ over $W(k)$ which, using (5.2.1) gives rise to a complex variety $A_{\mathbb{C}}$ over $\mathbb{C}$ (which depends only on the restriction $\varepsilon \mid W_0(\bar{k})$ (see [Deligne 1969, p. 239]), which in turn, is determined by $\varepsilon \mid W(k)$). Let $\pi \in \mathrm{Gal}(\bar{k}/k)$ denote the Frobenius. The corresponding morphism $\pi_{A/k}$ (which on the structure sheaf of A is given by the k-linear ring endomorphism $f \mapsto f^q$) lifts to an automorphism F_A on $T = T_A = H_1(A_{\mathbb{C}}, \mathbb{Z})$, and the pair (T_A, F_A) is a Deligne module.

Theorem 5.3 [Deligne 1969; Howe 1995]. *The association $A \to (T_A, F_A)$, determined by the embedding* (5.2.1), *induces an equivalence between the category of n-dimensional ordinary abelian varieties* (*resp. polarized abelian varieties*) *over $k = \mathbb{F}_q$ and the category of Deligne modules* (*resp. polarized Deligne modules*) *over k, of rank $2n$.* □

5.4. In [Goresky and Tai 2019], we define a real structure on a polarized Deligne module (T, F, ω) to be a group homomorphism $\tau : T \to T$ such that

$$\tau^2 = I, \quad \tau F \tau^{-1} = V, \quad \omega(\tau x, \tau y) = -\omega(x, y).$$

The involution τ is a characteristic p analog of complex conjugation. There are finitely many ("real") isomorphism classes of principally polarized Deligne modules (of dimension $2n$ over $k = \mathbb{F}_q$) with real structure and principal level N structure, and a formula for this number is given in [Goresky and Tai 2019]. There, we follow the method of Kottwitz [1990] and show that the number of isomorphism classes of principally polarized Deligne modules with real structure is finite and is given by an adèlic orbital integral.

5.5. In order to conceptualize the contribution at p to this formula it is convenient to define a *Deligne module at p* (over $\mathbb{F}_q$, of rank $2n$) to be a pair (T_p, F_p) where T_p is a free $\mathbb{Z}_p$ module of rank $2n$ and $F_p : T_p \to T_p$ is a semisimple endomorphism whose characteristic polynomial $\sum_{i=0}^{2n} a_i x^i$ is q-palindromic,[1] with middle coefficient a_n a p-adic unit, half of whose roots in $\bar{\mathbb{Q}}_p$ are p-adic units and half of which are divisible by p, such that there exists $V_p : T_p \to T_p$ with $F_p V_p = V_p F_p = q$. (This implies that if λ is a root then so is q/λ.) A polarization of (T_p, F_p) is a

[1] Meaning that $a_i = q^{n-i} a_{2n-i}$ for $0 \le i \le n-1$.

$\mathbb{Z}_p$-valued symplectic form ω_p such that $\omega(F_p x, y) = \omega(x, V_p y)$. (The "positivity condition" does not make sense in this setting.) A real structure τ_p on (T_p, F_p, ω_p) is a symplectic involution of T_p with multiplier -1 that exchanges F_p and V_p. If (T, F, ω, τ) is a (real, polarized) Deligne module then tensoring with $\mathbb{Z}_p$ gives a (real, polarized) Deligne module at p.

5.6. ***The Tate module.*** Let (T, F) be a Deligne module over $k = \mathbb{F}_{p^a}$. From this we define a $\mathrm{Gal}(\bar{k}/k)$ module, for $\ell \neq p$ a (rational) prime,

$$T_\ell(T) = T \otimes_{\mathbb{Z}} \mathbb{Z}_\ell$$

with Galois action determined by the rule that $\pi \in \mathrm{Gal}(\bar{k}/k)$ acts by $F \otimes 1$. A polarization and/or a real structure on (T, F) induces one on $T \otimes \mathbb{Z}_\ell$ in an obvious way.

Let $\ell \neq p$ be prime. If A is an ordinary abelian variety with Tate module $T_\ell(A)$ and Deligne module (T_A, F_A) then there is a natural isomorphism of $\mathrm{Gal}(\bar{k}/k)$ modules $T_\ell(A) \cong T_\ell(T_A) = T_A \otimes \mathbb{Z}_\ell$.

6. The Dieudonné module of an ordinary variety

6.1. Let A be an ordinary abelian variety over $k = \mathbb{F}_{p^a}$. Denote by $M(A) := M(A[p^\infty])$ the covariant Dieudonné module associated to the p-divisible group $A[p^\infty]$. In this section we explicitly construct this Dieudonné module $M(A)$ (and quasi-polarization) directly from the Deligne module (T_A, F_A) (and a polarization). In fact, the Dieudonné module $M(A)$ depends only on the associated Deligne module $(T_p = T_A \otimes \mathbb{Z}_p,\ F_p = F_A \otimes \mathbb{Z}_p)$ at p. Although this material is well known to experts, we require specific equations for these modules that do not appear to be in the literature.

Given a universal choice of involution $\bar{\tau}$ of the Witt vectors (as in Appendix A) we show, in Section 6.11, that a real structure on (T_p, F_p) determines a real structure on $M(A)$.

6.2. Let (T_p, F_p) be a Deligne module at p, over $k = \mathbb{F}_{p^a}$. The same argument as in [Deligne 1969] shows that the endomorphism F_p determines a unique decomposition

$$T_p \cong T' \oplus T'' \tag{6.2.1}$$

preserved by F_p and V_p, such that F_p is invertible on T' and is divisible by q on T''. In fact, the module $T' \otimes \overline{\mathbb{Q}}_p$ is the sum of the eigenspaces of F_p whose eigenvalues in $\overline{\mathbb{Q}}_p$ are p-adic units while $T'' \otimes \overline{\mathbb{Q}}_p$ is the sum of the eigenspaces of F_p whose eigenvalues are divisible by p. For $t = (t', t'') \in T_p$ set

$$A_q(t', t'') = (t', qt'') \quad \text{and} \quad A_p(t', t'') = (t', pt''). \tag{6.2.2}$$

Then $F_p A_q^{-1} = A_q^{-1} F_p : T_p \to T_p$ is an isomorphism. Extend F_p and σ to $T_p \otimes W(\bar{k})$ by $F_p(t \otimes w) = F_p(t) \otimes w$ and $\sigma(t \otimes w) = t \otimes \sigma(w)$.

6.3. ***The Dieudonné module of a Deligne module.*** For a Deligne module (T_p, F_p) at p as above, define the covariant Dieudonné module $M(T_p, F_p)$ (which we denote simply by $M(T_p)$) to be the $\mathrm{Gal}(\bar{k}/k)$-invariant submodule of $T_p \otimes W(\bar{k})$ where $\pi \in \mathrm{Gal}(\bar{k}/k)$ acts as

$$\pi(t \otimes w) = A_q^{-1} F_p(t) \otimes \sigma^a(w), \tag{6.3.1}$$

so to be explicit,

$$M(T_p) = \{x \in T_p \otimes W(\bar{k}) \mid F_p(x) = A_q \sigma^{-a}(x)\} \tag{6.3.2}$$

with actions $\mathcal{F}(t \otimes w) = pA_p^{-1}(t) \otimes \sigma(w)$ and $\mathcal{V}(t \otimes w) = A_p(t) \otimes \sigma^{-1}(w)$.

6.4. The mapping A_q preserves the splitting of T_p which gives a splitting $M(T_p) = M(T') \oplus M(T'')$. The operator $\mathcal{F}$ is σ-linear; it is invertible on $M(T'')$ and it is divisible by p on $M(T')$. If $\alpha \in M(T)$ then

$$F_p(\alpha) = \mathcal{V}^a(\alpha),$$

that is, the mapping F_p has been factored as $F_p = \mathcal{V}^a$. (The preceding paragraphs may be dualized so as to define the *contravariant* Dieudonné modules $N(T) = N(T') \oplus N(T'')$ corresponding to the splitting (6.3.2), in which case the mapping $\mathcal{F}_p$ is invertible on $N(T')$, divisible by p on $N(T'')$ and one has $F_p = \mathcal{F}^a$. Despite this confusion we use the covariant Dieudonné module because the equations are a bit simpler.)

Proposition 6.5. *Let (T_p, F_p) be a Deligne module at p with $\mathbb{Z}_p$-rank equal to $2n$. Then its Dieudonné module $M(T_p)$ is a free module over $W(k)$ whose $W(k)$-rank also equals $2n$ and in fact there exists a $W(k)$-basis of $M(T_p)$ whose elements also form a $W(\bar{k})$ basis of $T_p \otimes W(\bar{k})$.*

The proof will appear in Appendix B. The following lemma will be needed in the proof of Proposition 7.3.

Lemma 6.6. *Let (T_p, F_p) be a Deligne module at p. The operator $\sigma(t \otimes w) = t \otimes \sigma(w)$ on $T_p \otimes W(\bar{k})$ preserves the Dieudonné module $M(T_p) \subset T_p \otimes W(\bar{k})$. Suppose $\Lambda \subset M(T_p) \otimes \mathbb{Q}_p$ is a $W(k)$-lattice. Then the following statements are equivalent.*

(1) *The lattice Λ is preserved by $\mathcal{F}$ and $\mathcal{V}$.*

(2) $p\Lambda \subset \mathcal{F}\Lambda \subset \Lambda$.

(3) $p\Lambda \subset \mathcal{V}\Lambda \subset \Lambda$.

(4) $A_p^{-1}\mathcal{V}\Lambda = \Lambda$.

Such a lattice is also preserved by σ.

Proof. The equivalence of (1), (2), (3) is straightforward. (See also the related (4.4.3) when Λ is symplectic). Such a lattice Λ is also preserved by F_p, V_p so by the argument of [Deligne 1969] it decomposes as $\Lambda = \Lambda' \oplus \Lambda''$ with $\Lambda' = M_{\mathbb{Q}}(T_p)' \cap \Lambda$ and $\Lambda'' = M_{\mathbb{Q}}(T_p)'' \cap \Lambda$. Then $\mathcal{V} \mid \Lambda'$ is invertible: Since $F_p = \mathcal{V}^a$ is invertible on Λ' it follows that $\mathcal{V}$ is surjective on Λ', and it is injective because it is injective on $M_{\mathbb{Q}}(T_p)'$. Similarly $\mathcal{F} \mid \Lambda''$ is invertible which implies (4). Conversely, suppose that $A_p^{-1}\mathcal{V}\Lambda = \Lambda$. Then $\mathcal{V}\Lambda \subset A_p\Lambda \subset \Lambda$ and $\mathcal{F}\Lambda = p\mathcal{V}^{-1}\Lambda = (pA_p^{-1})\Lambda \subset \Lambda$. Finally, the action of $A_p^{-1}\mathcal{V}$ on $M(T_p)$ coincides with that of σ^{-1}, so (4) implies $\sigma\Lambda = \Lambda$. □

6.7. Let A/k be an ordinary abelian variety with Deligne module (T_A, F_A). The associated finite group scheme $A[p^r] = \ker(\cdot p^r)$ decomposes similarly into a sum $A'[p^r] \oplus A''[p^r]$ of an étale-local scheme and a local-étale scheme, with a corresponding decomposition of the associated p-divisible group, $A[p^\infty] = A' \oplus A''$. Over $W(\bar{k})$ the finite étale group scheme $A'[p^r]$ becomes constant so there is a canonical isomorphism

$$A'[p^r] \cong p^{-r}T'_A/T'_A. \tag{6.7.1}$$

Proposition 6.8. *The isomorphism $A'[p^r] \cong p^{-r}T'_A/T'_A$ induces an isomorphism of covariant Dieudonné modules*

$$M(A) \cong M(T_A \otimes \mathbb{Z}_p).$$

6.9. ***Proof of Proposition 6.8.*** The module $M(T_A \otimes \mathbb{Z}_p)$ was defined in Section 6.3, so we need to determine the Dieudonné module $M(A)$ of the abelian variety A. First let us show that

$$M(A') \cong (T'_A \otimes W(\bar{k}))^{\mathrm{Gal}}, \tag{6.9.1}$$

where the action of $\pi = \sigma^a \in \mathrm{Gal}$, of $\mathcal{F}$ and $\mathcal{V}$ is given by

$$\begin{aligned} \pi.(t' \otimes w) &= F_A(t') \otimes \sigma^a(w), \\ \mathcal{F}(t' \otimes w) &= pt' \otimes \sigma(w), \\ \mathcal{V}(t' \otimes w) &= t' \otimes \sigma^{-1}(w). \end{aligned} \tag{6.9.2}$$

From (6.7.1), over $W(\bar{k})$, the covariant Dieudonné module of the finite group scheme $A'[p^r]$ is:

$$\overline{M}(A'[p^r]) = (p^{-r}T'_A/T'_A) \otimes_{\mathbb{Z}} W(\bar{k}) \cong (T'_A/p^rT'_A) \otimes_{\mathbb{Z}} W(\bar{k}) \tag{6.9.3}$$

with $\mathcal{F}(t' \otimes w) = pt' \otimes \sigma(w)$; see [Demazure 1972, p. 68]. Then (see [Demazure 1972, p. 71] or [Chai et al. 2014, §B.3.5.9, p. 350]),

$$\overline{M}(A') = \varprojlim \overline{M}(A'[p^r]). \tag{6.9.4}$$

Therefore

$$\begin{aligned} M(A') &= (\varprojlim (T'_A/p^r T'_A) \otimes W(\bar{k}))^{\mathrm{Gal}} \\ &\cong \big(\varprojlim \big(T'_A \otimes W(\bar{k})/p^r (T'_A \otimes W(\bar{k}))\big)\big)^{\mathrm{Gal}} \\ &\cong (T'_A \otimes W(\bar{k}))^{\mathrm{Gal}} \end{aligned}$$

with (étale) Galois action

$$\pi(t' \otimes w) = \pi(t') \otimes \pi(w) = F_A(t') \otimes \sigma^a(w). \tag{6.9.5}$$

Next, using double duality, we will show that $M(A'') \cong (T''_A \otimes W(\bar{k}))^{\mathrm{Gal}}$ where

$$\begin{aligned} \pi(t'' \otimes w) &= q^{-1} F_A(t'') \otimes \sigma^a(w), \\ \mathcal{F}(t'' \otimes w) &= t'' \otimes \sigma(w), \\ \mathcal{V}(t'' \otimes w) &= pt'' \otimes \sigma^{-1}(w). \end{aligned} \tag{6.9.6}$$

Let B denote the ordinary abelian variety that is dual to A with Deligne module (T_B, F_B) and corresponding p-divisible groups B', B''. Then B' is dual to A'' (and vice versa), hence it follows from (6.9.1) (see also [Chai et al. 2014, §B.3.5.9], [Demazure 1972, p. 72] and [Howe 1995, Proposition 4.5]) that:

$$\begin{aligned} \overline{M}(B') &= T'_B \otimes_{\mathbb{Z}_p} W(\bar{k}),^2 \\ \overline{M}(A'') &= \mathrm{Hom}_{W(\bar{k})}(\overline{M}(B'), W(\bar{k})),^3 \\ T'_B &= \mathrm{Hom}_{\mathbb{Z}_p}(T''_A, \mathbb{Z}_p).^4 \end{aligned}$$

From this, we calculate that the isomorphism

$$\Psi : T''_A \otimes W(\bar{k}) \to \mathrm{Hom}_{W(\bar{k})}(\mathrm{Hom}_{\mathbb{Z}_p}(T''_A, \mathbb{Z}_p) \otimes W(\bar{k}), W(\bar{k})) = \overline{M}(A'')$$

defined by

$$\Psi_{t'' \otimes w}(\phi \otimes u) = \phi(t'').wu$$

(for $t'' \in T''_A$, for $\phi \in \mathrm{Hom}(T''_A, \mathbb{Z}_p)$ and for $w, u \in W(\bar{k})$) satisfies:

$$\begin{aligned} (\pi.\Psi_{t'' \otimes w})(\phi \otimes u) &= \sigma^a \Psi_{t'' \otimes w}(\pi_B^{-1}(\phi \otimes u)) \\ &= \sigma^a \Psi_{t'' \otimes w}(F_B^{-1}\phi \otimes \sigma^{-a} u) \\ &= \sigma^a ((F_B^{-1}\phi)(t'').w.\sigma^{-a} u) \\ &= \phi(V_A^{-1}(t'')).\sigma^a(w).u) = (\Psi_{V_A^{-1} t'' \otimes \sigma^a(w)})(\phi \otimes u). \end{aligned}$$

[2] $\pi(t' \otimes w) = F_B(t') \otimes \sigma^a(w)$, $\mathcal{F}(t' \otimes w) = pt' \otimes \sigma(w)$.
[3] $\pi_A \psi(m) = \sigma^a \psi(\pi_B^{-1}(m))$, $\mathcal{F}\psi(m) = \sigma\psi(\mathcal{V}(m))$.
[4] $F_B \phi(t') = \phi V_A(t')$.

Therefore $\pi(t'' \otimes w) = V_A^{-1}(t'') \otimes \sigma^a(w) = q^{-1} F_A(t'') \otimes \sigma^a(w)$. Similarly,

$$(\mathcal{F}.\Psi_{t''\otimes w})(\phi \otimes u) = \Psi_{t''\otimes\sigma(w)}(\phi \otimes u),$$

hence $\mathcal{F}(t'' \otimes w) = t'' \otimes \sigma(w)$, which proves (6.9.6). Since $M(A'') = (\overline{M}(A''))^{\mathrm{Gal}}$, this together with (6.9.1) verifies that $M(A)$ satisfies the condition in (6.3.2) (with T_p replaced by $T_A \otimes \mathbb{Z}_p$). □

Proposition 6.10. *Let (T_p, F_p) be a Deligne module at p. Let $\omega : T_p \times T_p \to \mathbb{Z}_p$ be a symplectic form such that $\omega(Fx, y) = \omega(x, Vy)$ for all $x, y \in T_p$. Extending scalars to $W(\bar{k})$ then restricting to the Dieudonné module $M(T_p) \subset T_p \otimes W(\bar{k})$ gives a quasi-polarization*

$$\omega_p : M(T_p) \times M(T_p) \to W(k)$$

of $M(T_p)$. If the original form ω is nondegenerate up to homothety then the same is true of the form ω_p, with the same homothety constant.

Proof. The proof is a direct computation using the decomposition $T_p \cong T' \oplus T''$. □

6.11. ***Real structures.*** Let (T_p, F_p) be a Deligne module at p, with a polarization $\omega : T_p \times T_p \to \mathbb{Z}_p$. Let ω_p denote the resulting quasi-polarization on the covariant Dieudonné module $M(T_p)$. Let $\tau : T_p \to T_p$ be a real structure on (T_p, F_p) that is compatible with the polarization ω. Unfortunately, the mapping τ does not induce an involution on the Dieudonné module $M(T_p)$ without making a further choice.

Following Appendix A, choose and fix, once and for all, a continuous $K(k)$-linear involution $\bar{\tau} : K(\bar{k}) \to K(\bar{k})$ that preserves $W(\bar{k})$, so that $\bar{\tau}\sigma^a(w) = \sigma^{-a}\bar{\tau}(w)$. Then the following construction provides a functor from the category of polarized Deligne modules with real structure to the category of quasi-polarized Dieudonné modules with real structure.

Proposition 6.12. *With (T_p, F_p, ω, τ) as above, the mapping*

$$\tau_p : T_p \otimes W(\bar{k}) \to T_p \otimes W(\bar{k})$$

defined by $\tau_p(x \otimes w) = \tau(x) \otimes \bar{\tau}(w)$ is continuous and $W(k)$-linear. It preserves the Dieudonné module $M(T_p)$ and it satisfies $\tau_p \mathcal{F}^a = \mathcal{V}^a \tau_p$ and

$$\omega_p(\tau_p x, \tau_p y) = -\omega_p(x, y) \quad \text{for all } x, y \in M(T_p). \tag{6.12.1}$$

Proof. The mapping τ takes T' to T'' (and vice versa) because it exchanges the eigenvalues of F and V. If $x' \otimes w \in T' \otimes W(\bar{k})$ then

$$\begin{aligned}\tau_p \pi.(x' \otimes w) &= \tau_p(F(x') \otimes \sigma^a(w)) = V\tau(x') \otimes \sigma^{-a}\bar{\tau}(w) \\ &= \pi^{-1}(\tau(x') \otimes \bar{\tau}(w)) = \pi^{-1}\tau_p(x' \otimes w)\end{aligned}$$

which shows that τ_p takes $M(T')$ to $M(T'')$ (and vice versa). Similarly,

$$\begin{aligned}\tau_p\mathcal{F}^a(x'\otimes w) &= \tau_p(x'\otimes q\sigma^a(w)) = \tau(x')\otimes q\sigma^{-a}\bar{\tau}(w)\\ &= \mathcal{V}^a(\tau(x')\otimes\bar{\tau}(w)) = \mathcal{V}^a\tau_p(x'\otimes w).\end{aligned}$$

Similar calculations apply to any element $x''\otimes w\in T''\otimes W(\bar{k})$.

We now wish to verify (6.12.1). Let $Y = T_p\otimes\mathbb{Q}$. It is possible to decompose $Y = Y_1\oplus\cdots\oplus Y_r$ into an orthogonal direct sum of simple $\mathbb{Q}_p[F]$ modules that are preserved by τ (see, for example, [Goresky and Tai 2019, Lemma 4.3]). This induces a similar ω_p-orthogonal decomposition of

$$M(Y) = M(T_p)\otimes_{W(k)} K(k)$$

into submodules $M_i = M(Y_i)$ over the rational Dieudonné ring

$$\mathcal{A}_{\mathbb{Q}} = \mathcal{A}\otimes K(k) = K(k)[F, V]/(\text{relations}),$$

each of which is preserved by τ_p. Since this is an orthogonal direct sum, it suffices to consider a single factor, that is, we may assume that (V_p, F_p) is a simple $\mathbb{Q}_p[F]$-module.

As in (6.2.1) the $\mathbb{Q}_p$ vector space Y decomposes, $Y = Y'\oplus Y''$ where the eigenvalues of $F\,|\,Y'$ are p-adic units and the eigenvalues of $F\,|\,Y''$ are divisible by p. Then the same holds for the eigenvalues of $\mathcal{F}^a$ on each of the factors of

$$M(Y) = M(Y')\oplus M(Y'').$$

Moreover, these factors are cyclic $\mathcal{F}^a$-modules and τ_p switches the two factors. It is possible to find a nonzero vector $y'\in M(Y')$ so that y' is $\mathcal{F}^a$-cyclic in $M(Y')$ and so that $y'' = \tau_p(y)$ is $\mathcal{F}^a$-cyclic in $M(Y'')$. It follows that $y = y'\oplus y''$ is a cyclic vector for $M(Y)$ which is fixed under τ_p, that is, $\tau_p(y) = y$. We obtain a basis of $M(Y)$:

$$y,\ \mathcal{F}^a y,\cdots,\ \mathcal{F}^{a(2n-1)}y.$$

The symplectic form ω_p is determined by its values $\omega_p(y, \mathcal{F}^{aj}y)$ for $1\le j\le 2n-1$. But

$$\begin{aligned}\omega_p(\tau_p y, \tau_p\mathcal{F}^{aj}y) &= \omega_p(y, \tau_p\mathcal{F}^{aj}\tau_p y) = q^j\omega_p(y, \mathcal{F}^{-aj}y)\\ &= q^j q^{-j}\omega_p(\mathcal{F}^{aj}y, y) = -\omega_p(y, \mathcal{F}^{aj}y).\end{aligned}$$ □

7. Comparing lattices in the ordinary case

7.1. A twisted orbital integral (4.4.2) "counts" (real, symplectic) lattices in a Dieudonné module while an untwisted orbital integral counts (real, symplectic) lattices in a Deligne module. In this section we show that such lattices are in natural one-to-one correspondence. Let (T_p, F_p, ω, τ) be a polarized Deligne module (at p) with a real structure. By [Goresky and Tai 2019, Proposition B.4] there exists an

isomorphism $\Phi : T_p \otimes \mathbb{Q}_p \to \mathbb{Q}_p^{2n}$ which takes ω to the standard involution ω_0 and takes τ to the standard involution τ_0. It takes F_p to some element $\gamma \in G(\mathbb{Q}_p)$ and it takes the decomposition (6.2.1) to a decomposition $\mathbb{Q}_p^{2n} = V' \oplus V''$ where γ is invertible on V' and is divisible by q on V''. It also takes the operator A_q of (6.2.2) to an element $\alpha_q \in G(\mathbb{Q})$ in the centralizer $Z_\gamma(\mathbb{Q})$ such that $\alpha_q|V' = I$ and $\alpha_q|V'' = qI$.

The mapping $\bar{\Phi} = \Phi \otimes K(\bar{k})$ is compatible with the action (see Lemma 6.6) of σ, that is, $\bar{\Phi}(t \otimes \sigma(w)) = \sigma\bar{\Phi}(t \otimes w)$, and it takes the rational Dieudonné module $M_{\mathbb{Q}}(T_p) = M(T_p) \otimes \mathbb{Q}_p$ to the $K(k)$-vector space (see Section 6.3)

$$\mathcal{J}_{\mathbb{Q}}(\gamma) = \{x \in K(\bar{k})^{2n} \mid \gamma x = \alpha_q \sigma^{-a}(x)\}.$$

In Corollary B.5 we construct a symplectic basis Ψ of $\mathcal{J}_{\mathbb{Q}}(\gamma)$ giving the diagram

$$\begin{array}{ccccc} T_p \otimes_{\mathbb{Z}} K(\bar{k}) & \xrightarrow{\bar{\Phi}} & K(\bar{k})^{2n} & \xleftarrow{\Psi \otimes K(\bar{k})} & K(\bar{k})^{2n} \\ \uparrow & & \uparrow & & \uparrow \\ M_{\mathbb{Q}}(T_p) & \xrightarrow{\cong} & \mathcal{J}_{\mathbb{Q}}(\gamma) & \xleftarrow{\Psi} & K(k)^{2n} \end{array} \tag{7.1.1}$$

The involution $\tau_p = \tau \otimes \bar{\tau}$ in the first column becomes $\bar{\tau}_0 = \tau_0 \otimes \bar{\tau}$ in the second and third columns. The mapping $\Psi \otimes K(\bar{k}) \in G(K(\bar{k}))$ satisfies $\widetilde{\Psi} = \bar{\tau}_0 \Psi \tau_0^{-1} = \Psi$. As in Sections 3.2 and 4.3, the operator $\mathcal{F}\sigma^{-1}$ (in the first column) on $M_{\mathbb{Q}}(T_p)$ becomes (in the third column) multiplication by $\delta \in G(K(k))$. Let $u_p = \Psi \alpha_p \Psi^{-1}$. Then $\delta\sigma(w) = \Psi^{-1} p \alpha_p^{-1} \sigma(\Psi w)$ so $\delta = p u_p^{-1} \Psi^{-1} \sigma(\Psi)$ and its norm

$$N(\delta) = \delta\sigma(\delta) \cdots \sigma^{a-1}(\delta) = \Psi^{-1} q \alpha_q^{-1} \sigma^a(\Psi) = \Psi^{-1} q \gamma^{-1} \Psi$$

is $G(K(\bar{k}))$-conjugate to $q\gamma^{-1}$. Similarly, the action of $\mathcal{V}\sigma$ becomes (in the third column) multiplication by $\eta = \Psi^{-1} \alpha_p \sigma^{-1}(\Psi)$ whose norm is stably conjugate to γ. Notations for these operators are summarized in Table 1.

$T \otimes \mathbb{Z}_p$	$T \otimes W(\bar{k}) \to$	$W(\bar{k})^{2n}$	$\leftarrow W(\bar{k})^{2n}$
	$M_{\mathbb{Q}}(T)$	$\mathcal{J}_{\mathbb{Q}}(\gamma)$	$K(k)^{2n}$
F_p	F_p	γ	$\Psi^{-1}\gamma\Psi$
A_p	A_p	α_p	u_p
	$\mathcal{F}$	$p\alpha_p^{-1}\sigma$	$\delta\sigma$
	$\mathcal{V}$	$\alpha_p\sigma^{-1}$	$p\sigma^{-1}\delta^{-1}$
ω	ω_p	ω_0	ω_0
τ	$\tau_p = \tau \otimes \bar{\tau}$	$\bar{\tau}_0 = \tau_0 \otimes \bar{\tau}$	$\bar{\tau}_0$

Table 1. Notations for corresponding operators.

7.2. For each $\mathbb{Z}_p$-lattice $L \subset T_p \otimes \mathbb{Q}_p$ that is preserved by F_p and V_p we obtain a $W(k)$-lattice

$$\Lambda = (L \otimes W(\bar{k}))^{\mathrm{Gal}(\bar{k}/k)} \subset M_{\mathbb{Q}}(T_p)$$

where the Galois action is given by $\pi.(t \otimes w) = FA_q^{-1}(t) \otimes \sigma^a(w)$ for $t \in L$ and $w \in W(\bar{k})$ and where $\mathcal{F}$ is given by $\mathcal{F}(t \otimes w) = pA_p^{-1}(t) \otimes \sigma(w)$ from (6.9.2) and (6.9.6).

Proposition 7.3. *Suppose $p \neq 2$. This association $L \mapsto \Lambda$ induces a one-to-one correspondence between*

(A) *the set of $\mathbb{Z}_p$-lattices $L \subset T_p \otimes \mathbb{Q}_p$, symplectic up to homothety, that are preserved by F_p, V_p and τ, and*

(B) *the set of $W(k)$-lattices $\Lambda \subset M_{\mathbb{Q}}(T)$, symplectic up to homothety, that are preserved by $\mathcal{F}$, $\mathcal{V}$, τ_p.*

The choice of basis Φ determines a one-to-one correspondence between (A) *and*

(C) *the set* $\{z \in H(\mathbb{Q}_p)/H(\mathbb{Z}_p) \mid z^{-1}\alpha_q^{-1}\gamma z \in G(\mathbb{Z}_p)\}$

with H as in (2.0.2). *The basis Ψ determines a one to one correspondence between* (B) *and*

(D) *the set* $\{w \in H(K(k))/H(W(k)) \mid w^{-1}p^{-1}u_p\delta\sigma(w) \in \Gamma_W\}$.

Conjugation by $\Psi \in \mathrm{Sp}_{2n}(K(\bar{k}))$ takes the centralizer $Z_\gamma(\mathbb{Q}_p) \subset H(\mathbb{Q}_p)$ isomorphically to the twisted centralizer

$$S_\delta(K(k)) = \{w \in H(K(k)) \mid z^{-1}\delta\sigma(z) = \delta\} \subset H(K(k)).$$

The correspondence (C) $\leftrightarrow$ (D) *is equivariant with respect to the action of these centralizers.*

Proof. Using the symplectic isomorphism Φ (and $\overline{\Phi}$) the set (A) may be identified with

(A′) the set of $\mathbb{Z}_p$-lattices $L \subset \mathbb{Q}_p^{2n}$, symplectic up to homothety (with respect to the standard symplectic form ω_0), preserved by the standard involution τ_0 and the mappings γ, $q\gamma^{-1}$.

Step 1. Let us show that (A′) $\leftrightarrow$ (C). As in [Deligne 1969], the special properties (Section 5.5) of γ determine a decomposition $\mathbb{Q}_p^{2n} = V' \oplus V''$ where γ is invertible on V' and is divisible by q on V''. Then $\alpha_q \mid V' = I$ and $\alpha_q \mid V'' = qI$. The same holds for any lattice $L \subset \mathbb{Q}_p^{2n} = L' \oplus L''$ that is preserved by γ and by $q\gamma^{-1}$. Such a lattice L is also preserved by $q\gamma^{-1}$ if and only if $\alpha_q^{-1}\gamma : L \to L$ is an isomorphism.

Write $L = gL_0$ for some $g \in G(\mathbb{Q}_p)$. If L is also preserved by the involution τ then $g^{-1}\tilde{g}L_0 = L_0$ (where $\tilde{g} = \tau_0 g \tau_0^{-1}$) so $g^{-1}\tilde{g}$ is a 1-cocycle for $H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(\mathbb{Z}_p))$,

which is trivial (by [Goresky and Tai 2019, Proposition B.4], and using the fact that $p \neq 2$). So there exists $h \in \mathrm{Sp}_{2n}(\mathbb{Z}_p)$ such that $h^{-1}\tilde{h} = g^{-1}\tilde{g}$, thus $L = zL_0$ where $z = gh^{-1} \in \mathrm{GL}_n^*(\mathbb{Q}_p)$. Therefore we have that $\alpha_q^{-1}\gamma z L_0 = zL_0$ so that $z^{-1}\alpha_q^{-1}\gamma z \in G(\mathbb{Z}_p)$. Replacing z by zt (for any $t \in H(\mathbb{Z}_p)$ gives the same lattice $L = ztL_0$. This proves (C).

The correspondence (B) $\to$ (D) is similar (compare Proposition 4.4). By Lemma 6.6, if a lattice $\Lambda \subset M_{\mathbb{Q}}(T)$ is preserved by $\mathcal{F}, \mathcal{V}$ then it splits $\Lambda = \Lambda' \oplus \Lambda''$; both factors are preserved by $\mathcal{F}, \mathcal{V}$; and $p^{-1}A_p\mathcal{F}(\Lambda) = \Lambda$. Translating this into the third column of Table 1, we have a $W(k)$-lattice, $w\Lambda_0 \subset K(k)^{2n}$ (where $\Lambda_0 = W(k)^{2n}$ is the standard lattice) such that $p^{-1}u_p\delta\sigma(w\Lambda_0) = w\Lambda_0$ or $w^{-1}p^{-1}u_p\delta\sigma(w) \in G(W(k))$, which is condition (D).

Step 2. Next, we claim the mapping $L \mapsto \overline{\Lambda} = L \otimes W(\bar{k})$ determines a correspondence between the set (A′) and

(A″) the set of $W(\bar{k})$-lattices $\overline{\Lambda} \subset K(\bar{k})^{2n}$, symplectic up to homothety, that are preserved by γ, $q\gamma^{-1}$, τ_0, and σ.

Given $\overline{\Lambda}$ from (A″) write $\overline{\Lambda} = \beta\overline{\Lambda}_0$ for some $\beta \in G(K(\bar{k}))$, where $\overline{\Lambda}_0 = W(\bar{k})^{2n}$ is the standard lattice. Then $\beta^{-1}\sigma(\beta) \in \mathrm{Sp}_{2n}(W(\bar{k})^{2n})$ is a 1-cocycle for the Galois cohomology $H^1(\mathrm{Gal}(\overline{\mathbb{F}}_p/\mathbb{F}_p), \mathrm{Sp}_{2n}(W(\cdot)))$, that is, the cohomology which forms an index set for the collection of all $\mathbb{F}_p$-isomorphism classes of $\mathbb{F}_p$-forms of nondegenerate skew symmetric bilinear forms on $W(\bar{k})^{2n}$, of which there is only one, by [Milnor and Husemoller 1973, §3.5]. So it is trivial, which implies that $\overline{\Lambda} = z\overline{\Lambda}_0$ for some $z \in G(\mathbb{Q}_p)$. (That is, $\beta^{-1}\sigma(\beta) = s^{-1}\sigma(s)$ for some $s \in G(W(\bar{k}))$; take $z = \beta s^{-1}$.)

The element $z^{-1}\alpha_q^{-1}\gamma z$ is in $G(W(\bar{k}))$ and it is fixed under σ so it lies in $G(\mathbb{Z}_p)$. This implies $\alpha_q^{-1}\gamma z L_0 = zL_0$, hence L is preserved by γ and by $q\gamma^{-1}$. Moreover, $\overline{\Lambda}^{\perp} = c\overline{\Lambda}$ where $c^{-1} \in \mathbb{Q}_p^{\times}$ is the multiplier of z, so the lattice $\overline{\Lambda}$ comes from the lattice $L = z\mathbb{Z}_p^{2n}$ and the homothety constant may be taken to lie in $\mathbb{Q}_p^{\times}$. Finally, since $\tau_0(\overline{\Lambda}) = \overline{\Lambda}$, the same argument as in Step 1 implies that z may be chosen to lie in $H(\mathbb{Q}_p)$, hence the lattice L is also preserved by τ_0.

Step 3. According to Section 6.3, the mapping $\overline{\Phi} : T_p \otimes K(\bar{k}) \to K(\bar{k})^{2n}$ takes the Dieudonné module $M(T_p) \otimes \mathbb{Q}_p$ to the module

$$\mathcal{J}_{\mathbb{Q}}(\gamma) = \{x \in K(\bar{k})^{2n} \mid \gamma x = \alpha_q \sigma^{-a}(x)\}$$

on which the mappings $\mathcal{F}, \mathcal{V}$ become the following (for which we use the same symbols): $\mathcal{F}(x) = p\alpha_p^{-1}\sigma(x)$ and $\mathcal{V}(x) = \alpha_p\sigma^{-1}(x)$. Consider

(B′) the set of $W(k)$-lattices $\Lambda \subset \mathcal{J}_{\mathbb{Q}}(\gamma)$, symplectic up to homothety, that are preserved by $\mathcal{F}, \mathcal{V}, \tau_0$.

We claim functors $(A'') \leftrightarrow (B')$ defined by

$$\overline{\Lambda} \mapsto \Lambda = \overline{\Lambda} \cap \mathcal{J}_{\mathbb{Q}}(\gamma)$$
$$\overline{\Lambda} = \Lambda \otimes W(\bar{k}) \hookleftarrow \Lambda$$

define a one-to-one correspondence between lattices $\overline{\Lambda}$ of (A'') and lattices Λ of (B').

Given $\overline{\Lambda}$ from the set (A''), the set $\Lambda = \overline{\Lambda} \cap \mathcal{J}_{\mathbb{Q}}(\gamma)$ is clearly preserved by $\mathcal{F}$, $\mathcal{V}$, τ_0, but we need to prove that it is a lattice. In fact, it is a free $W(k)$-module of maximal rank, which follows from the same proof (Appendix B) as that of Proposition 6.5.

On the other hand, given a lattice Λ from the set B' we obtain a lattice

$$\overline{\Lambda} = \Lambda \otimes W(\bar{k}) \subset K(\bar{k})^{2n}.$$

It is clearly preserved by F, V, τ_0. It follows from Lemma 6.6 that it is also preserved by σ, so it is in the set A''. We claim that $\overline{\Lambda} \cap (\mathcal{J}_{\mathbb{Q}}(\gamma))) = \Lambda$. Choose a $W(k)$-basis $b_1, b_2, \cdots, b_{2n} \in T_p \otimes K(\bar{k})$ of Λ. If $v = \sum_i s_i b_i \in \overline{\Lambda} \cap (\mathcal{J}_{\mathbb{Q}}(\gamma))$ with $s_i \in W(\bar{k})$ then

$$v = \sum_i s_i b_i = \gamma^{-1}\sigma^{-a}\alpha_q \sum_i s_i b_i = \sum_i \sigma^{-a}(s_i)\gamma^{-1}\alpha_q \sigma^{-a}(b_i) = \sum_i \sigma^{-a}(s_i) b_i$$

which implies that $s_i \in W(k)$. Therefore $v \in \Lambda$.

In fact every lattice in the set (A'') arises in this way: given $\overline{\Lambda}$ let $\Lambda = \overline{\Lambda} \cap \mathcal{J}_{\mathbb{Q}}(\gamma)$. Then Proposition 6.5 implies that Λ admits a $W(k)$ basis whose elements form a $W(\bar{k})$ basis of $\overline{\Lambda}$. So the inclusion $\Lambda \to \overline{\Lambda}$ induces an isomorphism $\Lambda \otimes W(\bar{k}) \cong \overline{\Lambda}$. This completes the verification of $(A'') \leftrightarrow (B')$.

Step 4. The correspondence between (B) and (B') is straightforward.

Step 5. Suppose $z \in Z_\gamma(\mathbb{Q}_p)$. Then z preserves the eigenspace decomposition $\mathbb{Q}_p^{2n} = V' \oplus V''$ so it commutes with α_p. Then $w = \Psi^{-1} z \Psi \in S_\delta$ because

$$w\delta\sigma(w)^{-1} = \Psi^{-1} p \alpha_p^{-1} \sigma(\Psi) = \delta.$$

Conversely if $w \in S_\delta(K(\bar{k}))$, applying the norm gives $wN(\delta)w^{-1} = N(\delta)$ so $z = \Psi w \Psi^{-1} \in Z_\gamma(K(\bar{k}))$. Moreover z commutes with α_p (as above). Substituting $\delta = \Psi^{-1} p \alpha_p^{-1} \sigma(\Psi)$ into the equation $w\delta\sigma(w)^{-1} = w$ gives $z\sigma(z)^{-1} = 1$, so $z \in Z_\gamma(\mathbb{Q}_p)$.

The equivariance statement in Proposition 7.3 is easily verified. □

7.4. As in Lemma 4.1, the theory of Smith normal form (or rational canonical form) gives a one-to-one correspondence between the set (A') and

(C') the set $\{g \in H(\mathbb{Q}_p)/H(\mathbb{Z}_p) \mid g^{-1}\gamma g \in \Gamma_p I_q \Gamma_p\}$,

where $\Gamma_p = G(\mathbb{Z}_p)$, and as in (4.4.1), an identification between (B') and

(D') the set $\{g \in H(K(k))/H(W(k)) \mid g^{-1}\delta\sigma(g) \in \Gamma_W A_p \Gamma_W\}$.

7.5. Using the same procedure (due to [Kottwitz 1990]) as in Sections 4.2 and 4.3, we may identify the set of isomorphism classes of principally polarized Deligne modules at p with real structure that are $\mathbb{Q}_p$-isogenous to (T_p, F_p, ω, τ) with the quotient

$$Y(T_p) = I(T_p)\backslash \mathcal{Y}(T_p),$$

where $\mathcal{Y}(T_p)$ denotes the set of $\mathbb{Z}_p$-lattices $L \subset T_p \otimes \mathbb{Q}_p$ that are symplectic up to homothety (that is, $L^\vee = cL$ for some $c \in \mathbb{Q}_p^\times$) and preserved by F_p, V_p, and τ (that is, the set (A) of Proposition 7.3), and where $I(T_p)$ denotes the group of self isogenies of (T_p, F_p, ω, τ).

So the correspondence (A) $\leftrightarrow$ (C) $\leftrightarrow$ (C)′ $\leftrightarrow$ (B) $\leftrightarrow$ (B)′ of Proposition 7.3 and 7.4 means that the number of such isomorphism classes

$$|Y(T_p)| = |Z_\gamma(\mathbb{Q}_p)\backslash \mathcal{Y}(T_p)|$$

is given by any of the integrals

$$\text{(C)} \qquad \int_{Z_\gamma(\mathbb{Q}_p)\backslash H(\mathbb{Q}_p)} \chi(z^{-1}\alpha_q^{-1}\gamma z)\, dz$$

$$\text{(C}') \qquad = \int_{Z_\gamma(\mathbb{Q}_p)\backslash H(\mathbb{Q}_p)} \kappa(g^{-1}\gamma g)\, dg$$

$$\text{(B)} \qquad = \int_{S_\delta(K(k))\backslash H(K(k))} \chi_W(w^{-1}p^{-1}u_p\delta\sigma(w))\, dw$$

$$\text{(B}') \qquad = \int_{S_\delta(K(k))\backslash H(K(k))} \kappa_W(g^{-1}\delta\sigma(g))\, dg$$

where χ is the characteristic function on $G(\mathbb{Q}_p)$ of $\Gamma_p = G(\mathbb{Z}_p)$, χ_W is the characteristic function of $G(W(k))$, κ is the characteristic function on $G(\mathbb{Q}_p)$ of $\Gamma_p I_q \Gamma_p$, κ_W is the characteristic function on $G(K(k))$ of $\Gamma_W I_p \Gamma_W$ and where $H = \mathrm{GL}_n^* \subset G$ (note that $\gamma, \delta \notin H$).

Appendix A: Involutions on the Witt vectors

A.1. Fix a finite field $\Bbbk$ of characteristic $p > 0$ having $q = p^a = |\Bbbk|$ elements. Fix an algebraic closure $\overline{\Bbbk}$ and let $W(\Bbbk)$, $W(\overline{\Bbbk})$ denote the ring of (infinite) Witt vectors. These are lattices within the corresponding fraction fields, $K(\Bbbk)$ and $K(\overline{\Bbbk})$. Let $W_0(\overline{\Bbbk})$ be the valuation ring in the maximal unramified extension $K_0(\overline{\Bbbk})$ of $\mathbb{Q}_p \subset K(\Bbbk)$. We may canonically identify $W(\overline{\Bbbk})$ with the completion of $W_0(\overline{\Bbbk})$. Denote by $\pi : \overline{\Bbbk} \to \overline{\Bbbk}$ the Frobenius $\pi(x) = x^q$. It has a unique lift, which we also denote by $\pi : W(\overline{\Bbbk}) \to W(\overline{\Bbbk})$, and the cyclic group $\langle \pi \rangle \cong \mathbb{Z}$ is dense in the Galois group $G = \mathrm{Gal}(K_0(\overline{\Bbbk})/K(\Bbbk)) \cong \mathrm{Gal}(\overline{\Bbbk}/\Bbbk)$. If $L \supset \Bbbk$ is a finite extension, for simplicity we write $\mathrm{Gal}(L/\Bbbk)$ in place of $\mathrm{Gal}(K(L)/K(\Bbbk))$ and we write $\mathrm{Trace}_{L/\Bbbk}$ for the trace $W(L) \to W(\Bbbk)$.

Proposition A.2. *There exists a continuous $W(\Bbbk)$-linear mapping $\bar{\tau} : W(\overline{\Bbbk}) \to W(\overline{\Bbbk})$ such that*:

(1) $\bar{\tau}^2 = I$.

(2) $\bar{\tau}\pi = \pi^{-1}\bar{\tau}$.

(3) *For any finite extension $E/\Bbbk$, the mapping $\bar{\tau}$ preserves $W(E) \subset W(\overline{\Bbbk})$.*

(4) *For any finite extension $L \supset E \supset \Bbbk$, the following diagrams commute*:

$$\begin{array}{ccc} W(L) & \xrightarrow[\bar{\tau}]{} & W(L) \\ \downarrow{\scriptstyle \mathrm{Trace}_{L/E}} & & \downarrow{\scriptstyle \mathrm{Trace}_{L/E}} \\ W(E) & \xrightarrow[\bar{\tau}]{} & W(E) \end{array} \qquad \begin{array}{ccc} W(L) & \xrightarrow[\bar{\tau}]{} & W(L) \\ \uparrow & & \uparrow \\ W(E) & \xrightarrow[\bar{\tau}]{} & W(E) \end{array}$$

Such an involution will be referred to as an *antialgebraic involution of the Witt vectors*.

Proof. Let $E \supset \Bbbk$ be a finite extension of degree r. Recall that an element $\theta_E \in W(E)$ is a *normal basis generator* if the collection $\theta_E, \pi\theta_E, \pi^2\theta_E, \cdots, \pi^{r-1}\theta_E$ forms a basis of the lattice $W(E)$ over $W(\Bbbk)$. By simplifying and extending the argument in [Lenstra 1985], P. Lundström [1999] showed that there exists a compatible collection $\{\theta_E\}$ of normal basis generators of $W(E)$ over $W(\Bbbk)$, where E varies over all finite extensions of $\Bbbk$, and where "compatible" means that $\mathrm{Trace}_{L/E}(\theta_L) = \theta_E$ for any finite extension $L \supset E \supset \Bbbk$. Let us fix, once and for all, such a collection of generators. This is equivalent to fixing a "normal basis generator" θ of the free rank one module

$$\varprojlim_E W(E)$$

over the group ring

$$W[\![G]\!] = \varprojlim_E W(\Bbbk)[\mathrm{Gal}(E/\Bbbk)].$$

For each finite extension $E/\Bbbk$, define $\tau_E : W(E) \to W(E)$ by

$$\tau_E\left(\sum_{i=0}^{r-1} a_i \pi^i \theta_E\right) := \sum_{i=0}^{r-1} a_i \pi^{-i} \theta_E = \sum_{i=0}^{r-1} a_i \pi^{r-i} \theta_E,$$

where $a_0, a_1, \cdots, a_{r-1} \in W(\Bbbk)$. Then $\tau_E^2 = I$ and $\tau_E \pi = \pi^{-1}\tau_E$. We refer to τ_E as an antialgebraic involution of $W(E)$. The mapping τ_E is an isometry (hence, continuous) because it takes units to units. To see this, suppose $v \in W(E)$ is a unit and set $\tau_E(v) = p^r u$ where $u \in W(E)$ is a unit. Then

$$v = \tau_E^2(v) = p^r \tau_E(u) \in p^r W(E)$$

is a unit, hence $r = 0$.

Next, we wish to show, for every finite extension $L \supset E \supset \Bbbk$, that $\tau_L | W(E) = \tau_E$ (so that τ_E is well defined) and that $\tau_E \circ \mathrm{Trace}_{L/E} = \mathrm{Trace}_{L/E} \circ \tau_L$. We have an exact sequence

$$1 \to \mathrm{Gal}(L/E) \to \mathrm{Gal}(L/\Bbbk) \to \mathrm{Gal}(E/\Bbbk) \to 1.$$

For each $h \in \mathrm{Gal}(E/\Bbbk)$ choose a lift $\hat{h} \in \mathrm{Gal}(L/\Bbbk)$ so that

$$\mathrm{Gal}(L/\Bbbk) = \{\hat{h}g : \ h \in \mathrm{Gal}(E/\Bbbk),\, g \in \mathrm{Gal}(L/E)\}.$$

Let $x = \sum_{h \in \mathrm{Gal}(E/\Bbbk)} a_h h\theta_E \in W(E)$ where $a_h \in W(\Bbbk)$. Then

$$x = \sum_{h \in \mathrm{Gal}(E/\Bbbk)} a_h h \sum_{g \in \mathrm{Gal}(L/E)} g\theta_L = \sum_{h \in \mathrm{Gal}(E/\Bbbk)} a_h \sum_{g \in \mathrm{Gal}(L/E)} \hat{h} g \theta_L$$

so that

$$\begin{aligned} \tau_L(x) &= \sum_{h \in \mathrm{Gal}(E/\Bbbk)} a_h \sum_{g \in \mathrm{Gal}(L/E)} \hat{h}^{-1} g^{-1} \theta_L \\ &= \sum_{h \in \mathrm{Gal}(E/\Bbbk)} a_h \hat{h}^{-1} \sum_{g \in \mathrm{Gal}(L/E)} g^{-1}\theta_L = \sum_{h \in \mathrm{Gal}(E/\Bbbk)} a_h h^{-1} \theta_E = \tau_E(x). \end{aligned}$$

To verify that $\tau_E \circ \mathrm{Trace}_{L/E}(x) = \mathrm{Trace}_{L/E} \circ \tau_L(x)$, it suffices to consider basis vectors $x = \hat{h} g \theta_L$ where $g \in \mathrm{Gal}(L/E)$ and $h \in \mathrm{Gal}(E/\Bbbk)$. Then $\mathrm{Trace}_{L/E}(x) = h\theta_E$ and

$$\begin{aligned} \mathrm{Trace}_{L/E}(\tau_L(x)) &= \sum_{y \in \mathrm{Gal}(L/E)} y \hat{h}^{-1} g^{-1} \theta_L = \hat{h}^{-1} \sum_{z \in \mathrm{Gal}(L/E)} z \theta_L \\ &= h^{-1} \,\mathrm{Trace}(\theta_L) = \tau_E \,\mathrm{Trace}_{L/E}\,(x). \end{aligned}$$

It follows that the collection of involutions $\{\tau_E\}$ determines an involution

$$\bar{\tau} : W_0(\overline{\Bbbk}) \to W_0(\overline{\Bbbk})$$

of the maximal unramified extension of $W(\Bbbk)$. It is a continuous isometry (so it takes units to units) and it satisfies the conditions (1)–(4). Therefore it extends uniquely and continuously to the completion $W(\overline{\Bbbk})$. □

Appendix B: Applications of Galois cohomology

B.1. Throughout this section let k be a finite field with an algebraic closure $\bar{k}$ with Galois group $\mathrm{Gal} = \mathrm{Gal}(\bar{k}/k)$. Let $W(k)$ be the ring of Witt vectors over k. A bilinear form on a free finite-dimensional $W(k)$ module V is (strongly) nondegenerate if it induces an isomorphism $V \to \mathrm{Hom}_{W(k)}(V, W(k))$. Let ω_0 be the standard symplectic form whose matrix is $J = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$. In this section we recall some standard facts and applications from Galois cohomology.

Proposition B.2. *The Galois cohomology set* $H^1(\mathrm{Gal}(\bar{k}/k), \mathrm{GL}_n(W(\bar{k})))$ *is trivial.*

Proof. The proof follows from [SGA 3_{III} 1970] Exp. XXIV, Prop. 8.1(ii) and [Grothendieck 1968] Thm. 11.7 and Remark 11.8.3 although it takes some work to translate these very general results of Grothendieck into this setting. □

Proposition B.3. *The Galois cohomology set* $H^1(\mathrm{Gal}(\bar{k}/k), \mathrm{Sp}_{2n}(W(\bar{k}))$ *is trivial.*

Proof. The proof also follows from [SGA 3_{III} 1970] and [Grothendieck 1968] but it also follows directly from Proposition B.2 as follows. There is a natural one-to-one correspondence between the set of $W(k)$-isomorphism classes of (strongly) nondegenerate alternating bilinear forms on $W(\bar{k})^{2n}$ and elements of

$$\ker\Big(H^1\big(\mathrm{Gal}(\bar{k}/k), \mathrm{Sp}_{2n}(W(\bar{k}))\big) \to H^1\big(\mathrm{Gal}(\bar{k}/k), \mathrm{GL}_{2n}(W(\bar{k}))\big)\Big).$$

In fact, if $\{\xi_\theta\}$ is a 1-cocycle (with $\theta \in \mathrm{Gal}$) which lies in this kernel then there exists $g \in \mathrm{GL}_{2n}(W(\bar{k}))$ so that $\xi_\theta = \theta(g)g^{-1}$ (for all $\theta \in \mathrm{Gal}$). It may be used to twist the standard symplectic form ω_0 to give a new symplectic form with matrix $B = {}^t g J g^{-1}$. Then $\theta(B) = B$ so it defines a symplectic form on $W(k)^{2n}$ which is nondegenerate over $K(k)$ and also over $W(\bar{k})$, which implies that it is nondegenerate over $W(k)$, i.e., strongly nondegenerate.

If R is a principal ideal domain, it is well known (see, for example, [Goresky and Tai 2019, Lemma B.2]) that all strongly nondegenerate symplectic forms on R^{2n} are isomorphic over R. It follows that the above kernel contains a single element. By Proposition B.2 above, this implies that $H^1(\mathrm{Sp}_{2n}(W(\bar{k}))$ is trivial. □

Proposition B.4. *Define an action of the group* $\langle\tau_0\rangle \cong \mathbb{Z}/(2)$ *on* $\mathrm{Sp}_{2n}(W(k))$ *where the nontrivial element acts as conjugation by* $\tau_0 = \left(\begin{smallmatrix} -I & 0 \\ 0 & I \end{smallmatrix}\right)$. *If* $\mathrm{char}(k) \neq 2$ *then the nonabelian cohomology set* $H^1(\langle\tau_0\rangle, \mathrm{Sp}_{2n}(W(k))$ *is trivial.*

Proof. This follows from the same method as [Goresky and Tai 2019, Propositions B.4 and D.2]: since $W(k)$ is a principal ideal domain containing $1/2$, every involution of $\mathrm{Sp}_{2n}(W(k))$ with multiplier equal to -1 is conjugate to the standard involution $\tilde{g} = \tau_0 g \tau_0^{-1}$. The above nonabelian cohomology set counts the number of conjugacy classes of such involutions. □

Corollary B.5. *Let* V *be a finite-dimensional free* $W(\bar{k})$ *module together with a semilinear action of* $\mathrm{Gal}(\bar{k}/k)$. *Let* V^{Gal} *be the* $W(k)$*-module of Galois invariant elements.*

(1) *The module* V^{Gal} *is free over* $W(k)$ *and there exists a* $W(k)$*-basis of* V^{Gal} *which is also a* $W(\bar{k})$*-basis of* V.

(2) *If* ω *is a (strongly nondegenerate)* $W(\bar{k})$*-valued symplectic form on* V *such that* $\omega(\theta x, \theta y) = \theta\omega(x, y)$ *for all* $\theta \in \mathrm{Gal}(\bar{k}/k)$ *then* ω *restricts to a strongly nondegenerate* $W(k)$*-valued symplectic form on* V^{Gal} *and there exists a symplectic* $W(k)$*-basis of* V^{Gal} *that is also a symplectic* $W(\bar{k})$*-basis of* V.

(3) *In addition to* (2), *if* $\mathrm{char}(k) \neq 2$, *if* $\tau_p : V \to V$ *is an involution such that* $\tau_p\theta = \theta^{-1}\tau_p$ *for all* $\theta \in \mathrm{Gal}(\bar{k}/k)$ *and* $\omega(\tau_p x, \tau_p y) = -\omega(x, y)$ *then* τ_p *restricts to an involution on* V^{Gal} *and the symplectic basis* $\{e_1, \cdots, e_n, e_1^*, \cdots, e_n^*\}$ *of* V^{Gal} *may be chosen so that* $\tau_p(e_i) = -e_i$ *and* $\tau_p(e_i^*) = e_i^*$.

Proof. For part (1), let $m = \mathrm{rank}(V)$. Since the conclusion holds in the case that $V = W(\bar{k})^m$ it suffices to show that there exists a $\mathrm{Gal}(\bar{k}/k)$-equivariant isomorphism $V \to W(\bar{k})^m$. Choose any $W(\bar{k})$ isomorphism $\phi : V \to W(\bar{k})^m$ where $m = \dim(V)$. Then $\theta \mapsto \theta(\phi)\phi^{-1} \in \mathrm{GL}_m(W(\bar{k}))$ is a 1-cocycle so it equals $\theta(B)B^{-1}$ for some $B \in \mathrm{GL}_m(W(\bar{k}))$ by Proposition B.2. It follows that the isomorphism

$$\phi' = B^{-1}\phi : V \to W(\bar{k})^m$$

is Galois equivariant.

For part (2), let $m = 2n$ in the preceding argument. The conclusions of the argument hold for the standard symplectic form ω_0 on $W(\bar{k})^{2n}$ so it suffices to construct a $\mathrm{Gal}(\bar{k}/k)$-equivariant symplectic isomorphism $V \to W(\bar{k})^{2n}$. The same argument works: choose the original isomorphism $\phi : V \to W(\bar{k})^{2n}$ so as to take the symplectic form ω to the standard symplectic form ω_0. The same argument (using Proposition B.3 this time) gives $B \in \mathrm{Sp}_{2n}(W(\bar{k}))$ so the resulting isomorphism $\phi' = B^{-1}\phi : V \to W(\bar{k})^m$ is equivariant and symplectic.

For part (3), first use (2) to obtain a symplectic isomorphism $\phi : V^{\mathrm{Gal}} \to W(k)^{2n}$. The conclusions of the argument hold for the standard involution τ_0 so it suffices to modify this isomorphism so as to be equivariant with respect to the involutions τ_p and τ_0. The same argument (using Proposition B.4 this time) also works: set $\tilde{\phi} = \tau_0\phi\tau_p^{-1}$. Then $\tilde{\phi}\phi^{-1} \in \mathrm{Sp}_{2n}(W(k))$ is a 1-cocycle for the action of $\langle\tau_0\rangle$ and since the cohomology vanishes, the mapping ϕ may be modified so as to become equivariant with respect to the involutions. □

Acknowledgements

We thank Gopal Prasad for providing us with references for the Galois cohomology proof of Proposition B.2. We are grateful to an anonymous referee for carefully reading this paper and for offering many valuable suggestions and corrections. We thank our copy editor Fintan Hegarty for his meticulous proofreading and help with the formatting of this paper. An earlier version of this paper is included in [Goresky and Tai 2017].

References

[Adler 1979] A. Adler, "Antiholomorphic involutions of analytic families of abelian varieties", *Trans. Amer. Math. Soc.* **254** (1979), 69–94. MR Zbl

[Andrianov 1987] A. N. Andrianov, *Quadratic forms and Hecke operators*, Grundlehren der Math. Wissenschaften **286**, Springer, 1987. MR Zbl

[Chai et al. 2014] C.-L. Chai, B. Conrad, and F. Oort, *Complex multiplication and lifting problems*, Math. Surveys and Monogr. **195**, Amer. Math. Soc., Providence, RI, 2014. MR Zbl

[Comessatti 1925] A. Comessatti, "Sulle varietà abeliane reali, I", *Ann. Mat. Pura Appl.* **2**:1 (1925), 67–106. MR Zbl

[Comessatti 1926] A. Comessatti, "Sulle varietà abeliane reali, II", *Ann. Mat. Pura Appl.* **3**:1 (1926), 27–71. MR Zbl

[Deligne 1969] P. Deligne, "Variétés abéliennes ordinaires sur un corps fini", *Invent. Math.* **8** (1969), 238–243. MR Zbl

[Demazure 1972] M. Demazure, *Lectures on p-divisible groups*, Lecture Notes in Math. **302**, Springer, 1972. MR Zbl

[Drinfeld 1976] V. G. Drinfeld, "Coverings of p-adic symmetric domains", *Funkcional. Anal. i Priložen.* **10**:2 (1976), 29–40. In Russian; translated in *Funct. Anal. Appl.* **10**:2 (1976), 107–115. MR

[Goren 2002] E. Z. Goren, *Lectures on Hilbert modular varieties and modular forms*, CRM Monogr. Series **14**, Amer. Math. Soc., Providence, RI, 2002. MR Zbl

[Goresky and Tai 2003a] M. Goresky and Y. S. Tai, "Anti-holomorphic multiplication and a real algebraic modular variety", *J. Differential Geom.* **65**:3 (2003), 513–560. MR Zbl

[Goresky and Tai 2003b] M. Goresky and Y. S. Tai, "The moduli space of real abelian varieties with level structure", *Compos. Math.* **139**:1 (2003), 1–27. MR Zbl

[Goresky and Tai 2017] M. Goresky and Y. S. Tai, "Real structures on ordinary abelian varieties", preprint, 2017. arXiv

[Goresky and Tai 2019] M. Goresky and Y. S. Tai, "Ordinary points mod p of $\mathrm{GL}_n(\mathbb{R})$-locally symmetric spaces", *Pacific J. Math* **303**:1 (2019), 165–215.

[Gross and Harris 1981] B. H. Gross and J. Harris, "Real algebraic curves", *Ann. Sci. École Norm. Sup.* (4) **14**:2 (1981), 157–182. MR Zbl

[Grothendieck 1968] A. Grothendieck, "Le groupe de Brauer, III: Exemples et compléments", pp. 88–188 in *Dix exposés sur la cohomologie des schémas*, edited by A. Grothendieck and N. H. Kuiper, Adv. Stud. Pure Math. **3**, North-Holland, Amsterdam, 1968. MR Zbl

[Howe 1995] E. W. Howe, "Principally polarized ordinary abelian varieties over finite fields", *Trans. Amer. Math. Soc.* **347**:7 (1995), 2361–2401. MR Zbl

[Katz 1981] N. Katz, "Serre–Tate local moduli", pp. 138–202 in *Surfaces algébriques* (Orsay, France, 1976–1978), edited by J. Giraud et al., Lecture Notes in Math. **868**, Springer, 1981. MR Zbl

[Kottwitz 1990] R. E. Kottwitz, "Shimura varieties and λ-adic representations", pp. 161–209 in *Automorphic forms, Shimura varieties, and L-functions, I* (Ann Arbor, MI, 1988), edited by L. Clozel and J. S. Milne, Perspect. Math. **10**, Academic Press, Boston, 1990. MR Zbl

[Kottwitz 1992] R. E. Kottwitz, "Points on some Shimura varieties over finite fields", *J. Amer. Math. Soc.* **5**:2 (1992), 373–444. MR Zbl

[Lenstra 1985] H. W. Lenstra, Jr., "A normal basis theorem for infinite Galois extensions", *Nederl. Akad. Wetensch. Indag. Math.* **47**:2 (1985), 221–228. MR Zbl

[Li and Oort 1998] K.-Z. Li and F. Oort, *Moduli of supersingular abelian varieties*, Lecture Notes in Math. **1680**, Springer, 1998. MR Zbl

[Lundström 1999] P. Lundström, "Normal bases for infinite Galois ring extensions", *Colloq. Math.* **79**:2 (1999), 235–240. MR Zbl

[Manin 1963] Y. I. Manin, "Theory of commutative formal groups over fields of finite characteristic", *Uspehi Mat. Nauk* **18**:6 (1963), 3–90. In Russian; translated in *Russ. Math. Surv.* **18** (1963), 1–83. MR Zbl

[Messing 1972] W. Messing, *The crystals associated to Barsotti–Tate groups: with applications to abelian schemes*, Lecture Notes in Math. **264**, Springer, 1972. MR Zbl

[Milne and Shih 1981] J. S. Milne and K.-y. Shih, "The action of complex conjugation on a Shimura variety", *Ann. of Math.* (2) **113**:3 (1981), 569–599. MR Zbl

[Milnor and Husemoller 1973] J. Milnor and D. Husemoller, *Symmetric bilinear forms*, Ergeb. Math. Grenzgeb. **73**, Springer, 1973. MR Zbl

[Moonen 2001] B. Moonen, "Group schemes with additional structures and Weyl group cosets", pp. 255–298 in *Moduli of abelian varieties* (Texel, Netherlands, 1999), edited by C. Faber et al., Progr. Math. **195**, Birkhäuser, Basel, 2001. MR Zbl

[Nori and Srinivas 1987] M. V. Nori and V. Srinivas, "Canonical liftings", (1987). Appendix to V. B. Mehta and V. Srinivas, "Varieties in positive characteristic with trivial tangent bundle", *Compos. Math.* **64**:2 (1987), 191–212. MR Zbl

[Oda 1969] T. Oda, "The first de Rham cohomology group and Dieudonné modules", *Ann. Sci. École Norm. Sup.* (4) **2**:1 (1969), 63–135. MR Zbl

[Oort 2001] F. Oort, "A stratification of a moduli space of abelian varieties", pp. 345–416 in *Moduli of abelian varieties* (Texel, Netherlands, 1999), edited by C. Faber et al., Progr. Math. **195**, Birkhäuser, Basel, 2001. MR Zbl

[Pink 2005] R. Pink, "Finite group schemes", lecture notes, ETH Zürich, 2005, available at https://tinyurl.com/pinkschemes.

[Seppälä and Silhol 1989] M. Seppälä and R. Silhol, "Moduli spaces for real algebraic curves and real abelian varieties", *Math. Z.* **201**:2 (1989), 151–165. MR Zbl

[SGA 3_{III} 1970] M. Demazure and A. Grothendieck, *Schémas en groupes, Tome III: Structure des schémas en groupes réductifs, Exposés XIX–XXVI* (Séminaire de Géométrie Algébrique du Bois Marie 1962–1964), Lecture Notes in Math. **153**, Springer, 1970. MR Zbl

[Shimura 1975] G. Shimura, "On the real points of an arithmetic quotient of a bounded symmetric domain", *Math. Ann.* **215** (1975), 135–164. MR Zbl

[Silhol 1982] R. Silhol, "Real abelian varieties and the theory of Comessatti", *Math. Z.* **181**:3 (1982), 345–364. MR Zbl

[Spence 1972] E. Spence, "m-symplectic matrices", *Trans. Amer. Math. Soc.* **170** (1972), 447–457. MR Zbl

Received August 22, 2018. Revised October 27, 2018.

MARK GORESKY
SCHOOL OF MATHEMATICS
INSTITUTE FOR ADVANCED STUDY
PRINCETON, NJ
UNITED STATES
goresky@ias.edu

YUNG SHENG TAI
DEPARTMENT OF MATHEMATICS
HAVERFORD COLLEGE
HAVERFORD, PA
UNITED STATES
ystai@comcast.net

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.243

SPECTRAHEDRAL REPRESENTATIONS OF PLANE HYPERBOLIC CURVES

MARIO KUMMER, SIMONE NALDI AND DANIEL PLAUMANN

We describe a new method for constructing a spectrahedral representation of the hyperbolicity region of a hyperbolic curve in the real projective plane. As a consequence, we show that if the curve is smooth and defined over the rational numbers, then there is a spectrahedral representation with rational matrices. This generalizes a classical construction for determinantal representations of plane curves due to Dixon and relies on the special properties of real hyperbolic curves that interlace the given curve.

Introduction

Determinantal representations of plane curves are a classical topic in algebraic geometry. Given a form f (i.e., a homogeneous polynomial) of degree d in three variables with complex coefficients and a general form g of degree $d-1$, there exists a $d \times d$ linear matrix $M = xA + yB + zC$ such that f is the determinant of M and g a principal minor of size $d-1$ (see for example [Dolgachev 2012, Chapter 4]). The matrix M can be chosen to be symmetric if g is a *contact curve*, which means that all intersection points between the curves defined by f and g have even multiplicity. The construction of M from f and g is due to Dixon [1902] (following Hesse's much earlier study of the case $d = 4$). We refer to this construction as the *Dixon process*.

For real curves, the most interesting case for us is that of *hyperbolic curves*. The smooth hyperbolic curves are precisely the curves whose real points contain a set of $\lfloor d/2 \rfloor$ nested ovals in the real projective plane (plus a pseudoline if d is odd). A form $f \in \mathbb{R}[x, y, z]$ is hyperbolic if and only if it possesses a real symmetric determinantal representation $f = \det(M)$ such that $M(e) = e_1A + e_2B + e_3C$ is (positive or negative) definite for some point $e \in \mathbb{P}^2(\mathbb{R})$. This is the *Helton–Vinnikov theorem*, which confirmed a conjecture by Peter Lax [Helton and Vinnikov 2007].

The Helton–Vinnikov theorem received a lot of attention in the context of semidefinite programming, which was also part of the original motivation: the set of

MSC2010: 14H50, 14P99, 52A10.

Keywords: real algebraic curves, determinantal representations, spectrahedra, linear matrix inequalities.

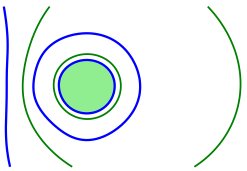


Figure 1. A quintic hyperbolic curve (blue), a quartic interlacer (green), and the hyperbolicity region (green region).

points $a \in \mathbb{R}^3$ for which the matrix $M(a)$ is positive semidefinite is a *spectrahedron*

$$\mathcal{S}(M) = \{a \in \mathbb{R}^3 : M(a) \succeq 0\}.$$

It coincides with the *hyperbolicity cone* $C(f, e)$ of $f = \det(M)$ in direction e, that is, the closure of the connected component of $\{a \in \mathbb{R}^3 : f(a) \neq 0\}$ containing e. This is a convex cone in $\mathbb{R}^3$, whose image in $\mathbb{P}^2$ is the region enclosed by the convex innermost oval of the curve (see Figure 1). A triple of real symmetric matrices A, B, C is a *spectrahedral representation* of $C(f, e)$ if $M = xA + yB + zC$ satisfies

$$C(f, e) = \mathcal{S}(M).$$

It has been pointed out by several authors [Vinnikov 2012; Plaumann and Vinzant 2013] that the proof of the Helton–Vinnikov theorem becomes much simpler if one requires the matrix M to be only hermitian, rather than real symmetric. In that case, M can be constructed via the Dixon process starting from any *interlacer* of f, that is, any hyperbolic form g of degree $d - 1$ whose ovals are nested between those of the curve defined by f (see Figure 1). One downside of this apparent simplification is that the corresponding determinantal representation $f = \det(M)$ with principal minor g is harder to construct explicitly, since one has to find the intersection points of f and g, while this can be avoided if g is a contact curve. We refer to [Vinnikov 2012] for a survey of these results.

In this paper, we study a modification of the Dixon process, which can be described as follows: given a form f of degree d, hyperbolic with respect to e, and an interlacer g of degree $d - 1$, we construct a real symmetric matrix pencil M with the properties that

- the determinant $\det(M)$ is divisible by f,
- the principal minor $\det(M_{11})$ is divisible by g,
- the extra factors $\det(M)/f$ and $\det(M_{11})/g$ are products of linear forms, and
- the spectrahedron defined by M coincides with $C(f, e)$.

The extra factor in our spectrahedral representation of $C(f, e)$ is an arrangement of real lines, as in Figure 2. Informally speaking, these additional lines correct the

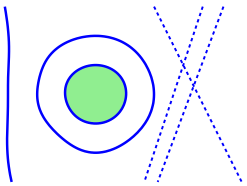


Figure 2. The extra factor (dashed blue lines) giving the spectrahedral representation of the hyperbolicity region.

failure of g to be a contact curve by passing through the intersection points of g with f that are not of even multiplicity.

The precise statement is Theorem 2.2. The size of M is at most quadratic in d. Thus, while M may not be the smallest or simplest determinantal representation of (some multiple of) f, it is easier to construct and may better reflect properties of the hyperbolicity region $C(f, e)$: as a corollary, we show that if f has coefficients in $\mathbb{Q}$, then $C(f, e)$ can be represented by a linear matrix inequality with coefficients in $\mathbb{Q}$ (Theorem 2.10). We may also view Theorem 2.2 in the context of the *generalized Lax conjecture*, which states that every hyperbolicity region (in any dimension) is spectrahedral. While various stronger forms of this conjecture have been disproved, it remains open as stated. One obstacle for constructing symmetric determinantal representations in higher dimensions is the nonexistence of contact interlacers for general hyperbolic hypersurfaces. Since our generalized Dixon process does not require the interlacer to be contact, it is possible that a spectrahedral description of the hyperbolicity cone could be constructed in a similar way, but this is currently purely speculative. In Section 3 we point out how our construction is related to sum-of-squares decompositions of Bézout matrices and the construction in [Kummer 2017].

Even in the original Dixon process for plane curves, details are somewhat subtle: for the construction to succeed as stated, the curve defined by f must be smooth, and the existence of a contact curve satisfying the required genericity assumption (equivalent to the existence of a nonvanishing even theta characteristic) was not rigorously established until somewhat later. Additionally, the case of singular curves was, to our knowledge, only fully settled and explicitly stated by Beauville [2000]. Likewise, in our generalized Dixon process, we need to treat degenerate cases with care and need some genericity assumptions.

Our generalized Dixon process has the additional feature that the size of the matrix M decreases if the interlacer g has real contact points with f. In particular, if g is an interlacer with *only real intersection points*, our statement reduces to that of the Helton–Vinnikov theorem. This leads us to the study of interlacers with real intersection (i.e., contact) points. Such interlacers are necessarily on the boundary of the cone $\mathrm{Int}(f, e)$ of all interlacers of f. An extreme ray of that cone

will necessarily have a certain number of real contact points (Lemma 1.3). However, we do not know whether there always exists an interlacer with the maximal number $d(d-1)/2$ of real contact points. Even in the case $d=4$, we only obtain a partial answer to this question (see the subsection beginning on page 248). There remain interesting (and easily stated) open questions concerning interlacing curves and the geometry of the interlacer cone.

1. Extremal interlacers

Let $f \in \mathbb{R}[x, y, z]$ be homogeneous of degree d and hyperbolic with respect to $e = (0:0:1)$, with $f(e) > 0$. Let $C = \mathcal{V}_{\mathbb{C}}(f)$ be the plane projective curve defined by f. We denote by $C(f, e)$ the closed hyperbolicity region of f with respect to e in the real projective plane.

Definition 1.1. Let $f, g \in \mathbb{R}[t]$ be univariate polynomials with only real zeros and with $\deg(g) = \deg(f) - 1$. Let $\alpha_1 \leq \cdots \leq \alpha_d$ be the roots of f, and let $\beta_1 \leq \cdots \leq \beta_{d-1}$ be the roots of g. We say that g *interlaces* f if $\alpha_i \leq \beta_i \leq \alpha_{i+1}$ holds for all $i = 1, \ldots, d-1$. If all these inequalities are strict, we say that g *strictly interlaces* f.

If $f \in \mathbb{R}[x, y, z]$ is hyperbolic with respect to e and g is homogeneous of degree $\deg(f) - 1$, we say that g *interlaces* f *with respect to* e if $g(te + v)$ interlaces $f(te + v)$ for every $v \in \mathbb{R}^3$. This implies that g is also hyperbolic with respect to e. We say that g *strictly interlaces* f if $g(te+v)$ strictly interlaces $f(te+v)$ for every $v \in \mathbb{R}^3$ not in $\mathbb{R}e$.

With f as above, let g be any form in $\mathbb{R}[x, y, z]$ coprime to f. We say that an intersection point $p \in \mathcal{V}_{\mathbb{C}}(f, g)$ is a *contact point* of g with f if the intersection multiplicity $\mathrm{mult}_p(f, g)$ is even. If all intersection points are contact points, then g is called a *contact curve* of f. A *curve of real contact* is a curve g for which all real intersection points are contact points, without any assumption on nonreal intersection points. Any interlacer is a curve of real contact.

Interlacers of f appear naturally in the context of determinantal representations of f [Plaumann and Vinzant 2013; Kummer et al. 2015]. For example, if $f = \det(xA+yB+zC)$ is a real symmetric and definite determinantal representation of f, then every principal $(d-1) \times (d-1)$ minor of $xA + yB + zC$ is an interlacer of f [Plaumann and Vinzant 2013, Theorem 3.3]. Furthermore, such a minor defines a contact curve (see, e.g., [Plaumann and Vinzant 2013, Proposition 3.2]). Conversely, given any interlacer of f that is also a contact curve, one can construct a definite determinantal representation of f and therefore a spectrahedral representation of its hyperbolicity region of size $d \times d$. However, for computational purposes, it is very difficult to actually find such an interlacer, even though its existence is guaranteed by the Helton–Vinnikov theorem [2007]. In Section 2, we will introduce

a method for constructing from an arbitrary interlacer a spectrahedral representation of possibly larger size. We denote by

$$\mathrm{Int}(f, e) = \{g \in \mathbb{R}[x, y, z]_{d-1} : g \text{ interlaces } f \text{ and } g(e) > 0\}$$

the set of interlacers of f. It is shown in [Kummer et al. 2015, Corollary 2.7] that this is a closed convex cone. Every boundary point of this cone has at least one contact point. In order to find interlacers with many contact points, it is therefore natural to consider extreme rays of this cone.

Definition 1.2. Let f be hyperbolic with respect to e. By an *extremal interlacer* of f we mean an extreme ray of the cone $\mathrm{Int}(f, e)$.

The next lemma gives a lower bound on the number of real contact points of an extremal interlacer.

Lemma 1.3. *Assume that f defines a smooth curve of degree d. Any extremal interlacer of f has at least*

$$\left\lceil \frac{(d+1)d-2}{4} \right\rceil$$

real contact points with f, counted with multiplicity.

Proof. Let g be an extremal interlacer, and let k be the number of real contact points of g. By definition, the real part of the divisor $\mathrm{div}_C(g)$ is even, say $2D$, with D real and effective of degree k. The space V of forms h of degree $d-1$ with $\mathrm{div}_C(h) \geq 2D$ has dimension at least $n = (d+1)d/2 - 2k$ and contains g. If $n > 1$, then V contains another form h linearly independent of g. We conclude that $g \pm \varepsilon h \in \mathrm{Int}(f, e)$ for sufficiently small ε. Thus, g is not extremal. Therefore, we must have $n \leq 1$, which gives $k \geq ((d+1)d-2)/4$. □

Remark 1.4. For smooth f, given any $d-1$ real points on the curve, there is an extremal interlacer touching the curve in (at least) the given points. Indeed, it is clear from the above proof that it suffices to show that there is an interlacer passing through these $d-1$ points. The quadratic system of interlacers considered in [Plaumann and Vinzant 2013, Definition 3.1] has dimension d, so we can prescribe $d-1$ points.

Remark 1.5. We do not know whether every hyperbolic curve possesses an *irreducible* extremal interlacer. This is true if C is a smooth cubic: for any two distinct points p and q on C, there is an extremal interlacing conic Q passing through p and q, by the preceding remark. If Q is reducible, it must factor into the two tangent lines to C at p and q. But Q is a contact curve by Lemma 1.3; hence, the intersection point of the two tangents must lie on C. Clearly, this will not be the case for a generic choice of p and q. This observation will be used at one point later on. It does not seem clear how to generalize this argument to higher degrees.

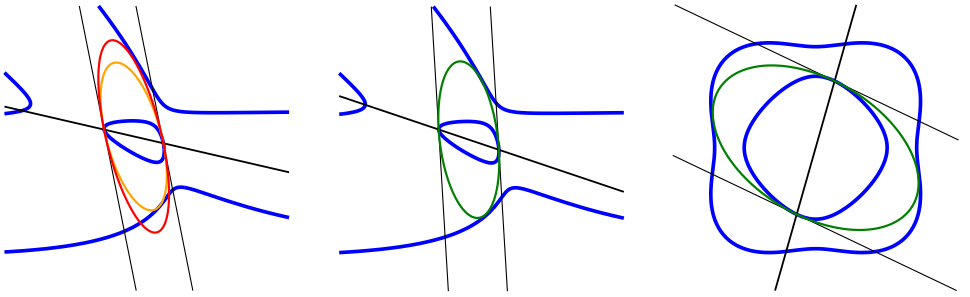


Figure 3. Quadrics touching hyperbolic quartics in real points.

The following table shows the expected number of real contact points of an extremal interlacer compared with the number of points for a full contact curve:

d	2	3	4	5	6	$\cdots$
$\lceil((d+1)d-2)/4\rceil$	1	3	5	7	10	$\cdots$
$d(d-1)/2$	1	3	6	10	15	$\cdots$

An interlacer can have many more real contact points than the estimate given by Lemma 1.3, and we do not know whether there is always one with only real intersection points.

Question 1.6. Does every hyperbolic plane curve have an interlacer that intersects the curve only in real points?

Even without the interlacing condition, it seems to be unknown whether a real curve always possesses a real contact curve with only real contact points. In the case of plane quartic curves we have some partial answers to that question.

The case of quartics. Let $C \subseteq \mathbb{P}^2$ be a smooth hyperbolic quartic that has a real bitangent touching C in only real points. We will show that in this case there is a contact interlacer touching C only in real points. It suffices to show that there is a conic touching both ovals in two real points. This, together with the above bitangent, will be the desired totally real interlacer.

Assume that $C(\mathbb{R})$ is contained in the affine chart $z \neq 0$ (for smooth quartic curves this is not a restriction). Let $l \in \mathbb{R}[x, y]_1$ be a nonzero linear form. Maximizing and minimizing l on the hyperbolicity region gives us two different linear polynomials l_1 and l_2 that are parallel and whose zero sets are tangent to the inner oval at some points p_1 and p_2 (see Figure 3).

Choose the signs such that both l_1 and l_2 are nonnegative on the inner oval. We consider the pencil of conics whose zero sets pass through p_1 and p_2 such that the tangent lines of the conics at p_1 and p_2 are defined by l_1 and l_2, respectively. This pencil is given by $q_\lambda = g^2 - \lambda l_1 l_2$, $\lambda \in \mathbb{R}$, where g is the line spanned by p_1 and p_2.

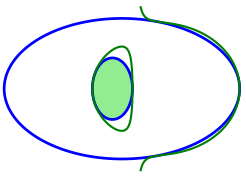


Figure 4. A hyperbolic quartic curve (in blue) and a cubic interlacer (in green) with only real intersection points.

The zero set of q_λ is completely contained in the interior of the outer oval for small $\lambda > 0$. Label the two half spaces defined by g by 1 and 2, and let $\lambda_i > 0$ be the smallest positive number such that the zero set of q_{λ_i} intersects the outer oval in the half space labeled by i. We observe that both q_{λ_i} have three real contact points with C. If $\lambda_1 = \lambda_2$, then we are done.

Now we let the linear form l, which we started with, vary continuously and we also keep track of the labels of the half spaces in a continuous manner. The resulting conic $q_{\lambda_1}(l)$ depends continuously on l, and we note that $q_{\lambda_1}(-l) = q_{\lambda_2}(l)$. Note that one of the zero sets of $q_{\lambda_1}(l)$ and of $q_{\lambda_1}(-l)$ on C contains a pair of complex conjugate points (the orange oval in Figure 3, left) whereas the other one contains only real points of C (the red oval in Figure 3, left). Therefore, there must be a linear form l_0 such that $q_{\lambda_1}(l_0)$ has the desired properties (Figure 3, center and right).

If there is no bitangent touching the quartic in two real points, we do not know whether there always exists an interlacer intersecting the curve in only real points. The next example shows that this is at least sometimes the case.

Example 1.7. We consider the smooth plane quartic defined by

$$f = 1250000x^4 - 1749500x^3y - 2250800x^2y^2 - 4312500x^2z^2 + 69260xy^3 + 786875xyz^2 + 88176y^4 + 1141000y^2z^2 + 1687500z^4.$$

Its real locus consists of two nested ovals both of which are convex (Figure 4), meaning that there is no bitangent touching the curve in two real points. Nevertheless, the interlacer given by

$$g = 500x^3 - 800x^2y - 740xy^2 - 625xz^2 + 176y^3 + 1000yz^2$$

intersects the quartic curve only in real points. Indeed, its divisor is given by

$$4 \cdot (4 : -5 : 0) + 2 \cdot (11 : 5 : 0) + 2 \cdot (1 : 5 : 0) + 2 \cdot (7 : 10 : -10) + 2 \cdot (7 : 10 : 10).$$

2. A generalized Dixon process

Given a real hyperbolic form f of degree d and an interlacer g of degree $d-1$, we wish to produce a real symmetric determinantal representation of f with a principal

minor divisible by g. If g is a contact curve, this is achieved through the classical Dixon process. We will extend the procedure in such a way that the resulting representation will reflect any real contact points between f and g, relating to our discussion of extremal curves of real contact in the previous section.

Let f be irreducible and hyperbolic with respect to $e \in \mathbb{P}^2(\mathbb{R})$, and assume that the plane curve $\mathcal{V}_{\mathbb{C}}(f)$ is smooth. Let g be an interlacer of f with r real contact points $p_1, \ldots, p_r$, counted with multiplicities. Consider the $d(d-1)-2r$ further intersection points, which are nonreal and therefore come in complex conjugate pairs, say $q_1, \ldots, q_s, \bar{q}_1, \ldots, \bar{q}_s$, so that $d(d-1) = 2r + 2s$. For each $i = 1, \ldots, s$ let ℓ_i be a linear form defining the unique (real) line joining q_i and $\overline{q_i}$. We will make the assumptions that

(G1) no three of the intersection points of f with g lie on a line,

(G2) no three of the ℓ_i pass through the same point, and

(G3) f does not vanish on any point where two of the ℓ_i intersect.

We begin by showing that such an interlacer always exists.

Lemma 2.1. *There exists a strict interlacer for which the genericity assumptions* (G1), (G2), *and* (G3) *are satisfied.*

Proof. Every choice of $k = \frac{1}{2}d(d+1) - 1$ points on the zero set of f that pose linearly independent conditions on forms of degree $d-1$ determines a unique such form. The other zeros of this $(d-1)$-form on the zero set of f depend continuously on the choice of the k points. By the general position theorem [Arbarello et al. 1985, Chapter III, §1], any neighborhood of the given interlacer contains a strict interlacer g with the property that its zero set intersects the one of f in $d(d-1)$ distinct points, any k of which pose linearly independent conditions on forms of degree $d-1$. Then we can slightly perturb any subset of k points in this intersection, and thus g, so that the number of triples of points in the intersection that lie on a line decreases. Thus, we can find a strict interlacer of f with the property that no three intersection points with the zero set of g lie on a line, so that genericity condition (G1) is satisfied. By the same argument, we can satisfy condition (G3).

For condition (G2), we need to move six points spanning three of the lines. Thus, the same argument applies, provided that $k \geq 6$, which means $d > 3$. The case $d \leq 2$ being trivial, we are left with condition (G2) for cubics ($d = 3$). In this case, we argue as follows. Suppose there is no interlacing conic satisfying condition (G2). Since the condition is Zariski-open, this would imply that condition (G2) is violated for any conic, strictly interlacing or not. But Lemma 1.3 and the subsequent Remark 1.5 imply that there exists an irreducible conic g touching f in three real points. Considering g as the limit of forms all of whose intersection points with f are simple, the assumption will imply that the three tangents to $\mathcal{V}(g)$ at the contact

points meet in one point. But since g is irreducible of degree 2, this is impossible. This contradiction shows the claim. □

Under these genericity assumptions, we will construct a symmetric linear determinantal representation M of $\ell_1 \cdots \ell_s \cdot f$ such that $\mathcal{S}(M)$ is the hyperbolicity region of f. Furthermore, the interlacer g divides a principal minor of M. The main result of this section is as follows.

Theorem 2.2. *Let f be an irreducible form of degree d that is hyperbolic with respect to $e \in \mathbb{P}^2(\mathbb{R})$, and assume that the plane curve $\mathcal{V}(f)$ is smooth. Let g be an interlacer of f with r real contact points, counted with multiplicities, that satisfies the genericity assumptions* (G1), (G2), *and* (G3). *Then there exists a symmetric linear matrix pencil M of size*

$$m = \frac{d^2 + d - 2r}{2}$$

which is positive definite at e and such that $C(f, e) = \mathcal{S}(M)$. We can choose M in such a way that g divides the principal minor $M_{1,1}$ of M and $\det(M)/f$ is a product of $m - d$ linear forms. Furthermore, each $(m-1) \times (m-1)$ minor M_{1l}, $1 \le l \le m$, of M is also divisible by the product of these $m - d$ linear forms.

The proof will consist of an algorithm that produces the desired representation given f and g.

We begin with some preliminaries. Given any two real ternary forms f, g of degrees d and d', respectively, without common components, we denote by $(f.g)$ the *intersection cycle* of f and g, consisting of the intersection points of the curves $\mathcal{V}(f)$ and $\mathcal{V}(g)$ in $\mathbb{P}^2(\mathbb{C})$. It is a 0-cycle, i.e., an element of the free abelian group over the points of $\mathbb{P}^2(\mathbb{C})$. Explicitly, $(f.g) = \sum_{i=1}^{k} m_i p_i$, with $\mathcal{V}(f) \cap \mathcal{V}(g) = \{p_1, \ldots, p_k\}$ and m_i positive integers, the intersection multiplicities. By Bézout's theorem, we have $\sum_{i=1}^{k} m_i = dd'$. Intersection cycles are additive, i.e., $((f_1 \cdot f_2).g) = (f_1.g) + (f_2.g)$. Furthermore, there is a natural partial order on 0-cycles, by comparing coefficients. We need the following classical result from the theory of plane curves, which we restate in the form we require.

Theorem 2.3 (Max Noether). *Let f, g, h be real ternary forms. Assume that f is irreducible and does not divide gh, and that the curve $\mathcal{V}(f) \subset \mathbb{P}^2(\mathbb{C})$ is smooth. If $(h.f) \ge (g.f)$, then there exist real forms a and b such that*

$$h = af + bg.$$

Proof. See [Fulton 1989, §5.5, Proposition 1]. □

Now let f and g be given as in the statement of Theorem 2.2, with intersection points $p_1, \ldots, p_r, q_1, \ldots, q_s, \overline{q_1}, \ldots, \overline{q_s}$ as before, and let ℓ_i be the linear form

defining the line between q_i and $\overline{q}_i$, for $i = 1, \ldots, s$, under the genericity assumptions (G1)–(G3).

Put $h = \ell_1 \cdots \ell_s$, and consider the polynomial fh of degree $(d^2+d-2r)/2 = m$, which is hyperbolic with respect to e. Furthermore, since each line ℓ_i meets C in the nonreal point q_i, none of the lines pass through $C(f, e)$, so that $C(fh, e) = C(f, e)$.

It therefore suffices to construct a symmetric linear determinantal representation of fh which is definite at e. This can be carried out with a modification of Dixon's method, which we now describe in several steps.

(1) Let V be the linear space of real forms of degree $d-1$ vanishing at $p_1, \ldots, p_r$. We have $\dim(V) \geq (d+1)d/2 - r = d + s$, and we pick linearly independent forms $a_1, \ldots, a_{d+s} \in V$, with $a_1 = g$. We introduce names for all the occurring intersection points:

$$
\begin{aligned}
(a_1.f) &= (g.f) = 2\sum_{j=1}^{r} p_j + \sum_{j=1}^{s}(q_j + \overline{q}_j), \\
(a_i.f) &= \sum_{j=1}^{r} p_j + \sum_{j=1}^{r+2s} p_{ij} \qquad \text{for } i \geq 2, \\
(\ell_i.f) &= q_i + \overline{q}_i + \sum_{j=1}^{d-2} r_{ij}, \\
(\ell_i.\ell_j) &= s_{ij} \qquad\qquad\qquad\quad \text{for } i \neq j.
\end{aligned}
$$

(2) Fix $k, l \in \{2, \ldots, d+s\}$ with $k \leq l$. We wish to find a real form b_{kl} of degree $d+s-1$ such that

$$
b_{kl} g - h a_k a_l \in (f) \tag{2.4}
$$

by applying Max Noether's theorem: we compute the intersection cycles

$$
\begin{aligned}
(ha_k a_l.f) &= 2\sum_{j=1}^{r} p_j + \sum_{j=1}^{s}(q_j + \overline{q}_j) + \sum_{j=1}^{s}\sum_{j'=1}^{d-2} r_{jj'} + \sum_{j=1}^{r+2s} p_{kj} + \sum_{j=1}^{r+2s} p_{lj}, \\
(g.f) &= 2\sum_{j=1}^{r} p_j + \sum_{j=1}^{s}(q_j + \overline{q}_j)
\end{aligned}
$$

and thus find b_{kl} with

$$
(b_{kl}.f) = \sum_{j=1}^{s}\sum_{j'=1}^{d-2} r_{jj'} + \sum_{j=1}^{r+2s} p_{kj} + \sum_{j=1}^{r+2s} p_{lj}.
$$

(3) Assume that $k = l$. Then we will produce a real form q of degree $s-1$ such that $c_{kk} := b_{kk} + qf$ satisfies

$$(c_{kk}.\ell_i) = (b + qf.\ell_i) = \sum_{j=1}^{d-2} r_{ij} + \sum_{j\neq i} s_{ij} + 2t_{ki}$$

for some real point $t_{ki} \in \ell_{ki}$, for all $i = 1, \ldots, s$. To this end, we let ℓ_0 be a linear form which does not vanish on any of the s_{ij}. Let $h_{ij} = (\ell_0 \cdots \ell_s)/(\ell_i \ell_j)$ and $\alpha_{ij} = -b_{kk}(s_{ij})/(h_{ij}(s_{ij}) f(s_{ij}))$ for $1 \leq i < j \leq s$. Note that h_{ij} vanishes on all s_{mn} except for s_{ij}. After replacing b_{kk} by $b_{kk} + \sum_{i,j} \alpha_{ij} h_{ij} f$, we can thus assume that b_{kk} vanishes on all the s_{ij}.

Next, we consider

$$q_\alpha = \sum_{j=1}^{s} \alpha_j \frac{\ell_1 \cdots \ell_s}{\ell_j}$$

with $\alpha_1, \ldots, \alpha_s \in \mathbb{R}$. The form q_α satisfies $q_\alpha(s_{ij}) = 0$ for all $j \neq i$ for any choice of the α_j. If we now take $q = \tilde{q} + q_\alpha$, we find

$$(b_{kk} + qf.\ell_i) = \sum_{j=1}^{d-2} r_{ij} + \sum_{j\neq i} s_{ij} + u_i + v_i$$

with u_i and v_i depending on α. Restricting to ℓ_i we therefore get $b_{kk} + qf = P \cdot (\tilde{b} + \alpha_i \tilde{f})$ where P is a nonzero polynomial whose roots are the r_{ij} and s_{ij}, and where $\tilde{b}$ and $\tilde{f}$ are polynomials of degree two. After possibly replacing α_i by its negative, we can assume that $\tilde{f}$ is strictly positive on ℓ_i since it has no real zeros on ℓ_i. Therefore, we can choose α_i in such a way that $\tilde{b} + \alpha_i \tilde{f}$ has a double zero t_{ki} and that makes the product of $b_{kk} + qf$ and $f \cdot ((\ell_1 \cdots \ell_s)/\ell_i)\ell_i(e)$ nonnegative on ℓ_i. The reasons for the latter requirement will become clear in a later step.

(4) Similarly, if $k < l$, we can find a real form q of degree $s-1$ such that $c_{kl} := b_{kl} + qf$ satisfies

$$(c_{kl}.\ell_i) = (b_{kl} + qf.\ell_i) = \sum_{j=1}^{d-2} r_{ij} + \sum_{j\neq i} s_{ij} + t_{ki} + t'_{ki}$$

for some real point $t'_{ki} \in \ell_i$. In fact, we even have that $t'_{ki} = t_{li}$. Indeed, this follows from (2.4) and the following lemma applied to each ℓ_i.

Lemma 2.5. *Let $f \in \mathbb{R}[t]$ be a polynomial of degree two without real zeros. Let $a, b, c \in \mathbb{R}[t]$ be polynomials of degree at most two such that a and c both have a double zero, ac is nonnegative, and b vanishes at the zero of a. If $ac = b^2 \bmod f$, then b vanishes at the zero of c as well.*

Proof. Let $a = \alpha(t-\beta)^2$, $c = \alpha'(t-\beta')^2$, and $b = \gamma(t-\beta)(t-\beta'')$ for some $\alpha, \alpha', \beta, \beta', \beta'', \gamma \in \mathbb{R}$ with $\alpha\alpha' \geq 0$. We have by assumption

$$\alpha\alpha'(t-\beta)^2(t-\beta')^2 = \gamma^2(t-\beta)^2(t-\beta'')^2 \bmod f.$$

Since $\mathbb{R}[t]/(f)$ is isomorphic to the field of complex numbers, it follows that

$$\alpha\alpha'(t-\beta')^2 = \gamma^2(t-\beta'')^2 \bmod f.$$

If $\gamma \neq 0$, then $\alpha\alpha' > 0$ and $t-\beta' = \pm\sqrt{\gamma^2/(\alpha\alpha')}\cdot(t-\beta'') \bmod f$. Finally, it follows that $\alpha\alpha' = \gamma$ and that $\beta' = \beta''$ because $1, t \in \mathbb{R}[t]/(f)$ are $\mathbb{R}$-linearly independent. □

If $k > l$, we let $c_{kl} = c_{lk}$.

(5) We now put $c_{1k} = c_{k1} = ha_k$ and consider the matrix N with entries c_{kl}, for $k, l = 1, \ldots, d+s$. By construction, the (2×2)-minors

$$c_{11}c_{kl} - c_{1k}c_{1l} = hgc_{kl} - h^2a_ka_l = h(gc_{kl} - ha_ka_l)$$

are divisible by fh. Since the first row of N is not divisible by f, it follows that all (2×2)-minors of N are divisible by f. We need to show that all (2×2)-minors $c_{kl}c_{k'l'} - c_{kl'}c_{k'l}$ are also divisible by h. Let u be such a minor, and fix $i \in \{1, \ldots, s\}$. Note that u has degree $2d+2s-2$ and vanishes (with multiplicities) on the $2d+2s-2$ points $2\sum_{j=1}^{d-2} r_{ij}$, $2\sum_{j\neq i} s_{ij}$, and $(t_{ki} + t_{k'i} + t_{li} + t_{l'i})$ on ℓ_i, since both products $c_{kl}c_{k'l'}$ and $c_{kl'}c_{k'l}$ vanish at those points. Since u is divisible by f, it also vanishes at $q_i + \overline{q}_i$. Thus, u vanishes identically on ℓ_i for each i, which implies $h \mid u$.

(6) In this step we show that c_{22} interlaces fh. This can be done by proving that $c_{22}\cdot \mathrm{D}_e(fh)$ is nonnegative on the zero set of fh [Kummer et al. 2015, Theorem 2.1]. Here $\mathrm{D}_e(fh)$ denotes the derivative of fh in direction e. We have

$$\mathrm{D}_e(fh) = h\cdot \mathrm{D}_e f + f\cdot \sum_{i=1}^{s} \ell_i(e)\frac{\ell_1\cdots\ell_s}{\ell_i}.$$

We can rewrite this modulo f and find

$$c_{22}\cdot \mathrm{D}_e(fh) = c_{22}\cdot h\cdot \mathrm{D}_e f = \frac{ha_2^2}{g}\cdot h\cdot \mathrm{D}_e f = \frac{\mathrm{D}_e f}{g}h^2a_2^2 \bmod f$$

by (2.4). This is nonnegative on the zero set of f because both $\mathrm{D}_e f$ and g are interlacers. On the other hand, modulo ℓ_i we obtain

$$c_{22}\cdot \mathrm{D}_e(fh) = c_{22}\cdot \ell_i(e)\cdot \frac{\ell_1\cdots\ell_s}{\ell_i} \bmod \ell_i,$$

which is nonnegative on the line defined by ℓ_i by the choices made in step (3).

(7) Now we proceed as in the usual Dixon process, referring to [Plaumann and Vinzant 2013] for details. Since all (2×2)-minors of the $(d+s)\times(d+s)$-matrix N

are divisible by fh, its maximal minors are divisible by $(fh)^{d+s-2}$ (see for example [Plaumann and Vinzant 2013, Lemma 4.7]). The signed maximal minors of N have degree $(d+s-1)^2$ and are the entries of the adjugate matrix N^{adj}. It follows that

$$M = (fh)^{2-d-s} \cdot N^{\mathrm{adj}}$$

has linear entries. Using the familiar identity $NN^{\mathrm{adj}} = \det(N) \cdot I_{d+s}$, we conclude

$$\det(M) = \gamma \cdot fh$$

for some constant $\gamma \in \mathbb{R}$. It remains to show that $\gamma \neq 0$. Suppose $\gamma = 0$; then $\det(M)$ is identically zero and hence so is $\det(N)$. In particular, the matrix $N(e)$ is singular. Let $\lambda \in \mathbb{R}^{d+s}$ be a nontrivial vector in the kernel of $N(e)$, and consider the polynomial $\tilde{g} = \lambda^t N \lambda$. It follows from the linear independence of the entries of the first row of N that $\tilde{g}$ is not the zero polynomial [Plaumann and Vinzant 2013, Lemma 4.8]. Since c_{22} interlaces fh by (6), so does $\tilde{g}$ [Plaumann and Vinzant 2013, Theorem 3.3, (1) $\Longrightarrow$ (2)], contradicting $\tilde{g}(e) = 0$. That $M(e)$ is definite also follows from the fact that c_{22} interlaces fh, by [Plaumann and Vinzant 2013, Theorem 3.3, (2) $\Longrightarrow$ (3)]. Note that the result in [Plaumann and Vinzant 2013] is stated only for irreducible curves. However, the same argument will apply here, since c_{22} is coprime to fh (unlike c_{11}, which is divisible by h). Indeed, we have chosen c_{22} in step (3) in such a way that it does not vanish entirely on any of the lines l_i. Thus, c_{22} is coprime to h. Moreover, c_{22} is congruent to b_{22} modulo (f). Thus, if f divided c_{22}, it would also divide a_2 by (2.4), which is not the case.

This finishes the construction of the determinantal representation M of fh. Finally, we note that the spectrahedron $\mathcal{S}(M)$ coincides with the hyperbolicity region $C(f, e)$ of f. Since $\det(M) = f \cdot \ell_1 \cdots \ell_s$, this simply amounts to the fact that the lines $\ell_1, \ldots, \ell_s$ do not pass through $C(f, e)$. Indeed, each ℓ_j has two nonreal intersection points with C, while lines passing through the hyperbolicity region will meet C in only real points. This completes the proof of Theorem 2.2.

Remark 2.6. Clearly, the corank of the constructed matrix pencil M is at least one at each point where fh vanishes. It can have corank more than one only at singularities of fh, i.e., in our case the points where two components intersect. Since the adjugate $N = M^{\mathrm{adj}}$ vanishes identically at the points r_{ij} and s_{ij} and because these are ordinary nodes, the corank of M at these points is exactly two. On the other hand, we have constructed N in such a way that it is not entirely zero at the points q_j and $\overline{q_j}$. Thus, M has corank one at these points. This shows in particular that M is not equivalent to a block diagonal matrix with more than one block.

Remark 2.7. The vector space V in step (1) of our construction can be found without computing all the real contact points $p_1, \ldots, p_r$. Indeed, by genericity assumption (G1) the $q_i, \overline{q_i}$ are all simple intersection points. Therefore, the p_i can

be computed as the singular locus of the zero-dimensional scheme cut out by f and g via the Jacobian criterion.

Next we observe that the genericity assumption in the theorem, as well as the smoothness assumption on f, can be dropped for strict interlacers by applying a limit argument.

Corollary 2.8. *Let f be a real form of degree d that is hyperbolic with respect to $e \in \mathbb{P}^2(\mathbb{R})$, and let g be a strict interlacer of f. Then there exists a symmetric linear matrix pencil M of size $(d^2+d)/2$ which is definite at e and such that $C(f,e) = \mathcal{S}(M)$. We can choose M in such a way that g divides a principal minor of M and $\det(M)/f$ is a product of $(d^2-d)/2$ linear forms.*

Proof. Let $m = (d^2+d)/2$. We may assume that $f(e) = 1$ and consider only monic representations $f = \det(M)$, i.e., with $M(e) = I_m$. The determinant map taking a monic symmetric real linear matrix pencil of size $m \times m$ to its determinant is proper; hence, its image is closed (see for example [Plaumann and Vinzant 2013, Lemma 3.4]). If g is a strict interlacer of f, the pair (f, g) is in the closure of the set of pairs $(\tilde{f}, \tilde{g})$, where $\tilde{f}$ is hyperbolic with respect to e, $\mathcal{V}(\tilde{f})$ is smooth, and $\tilde{g}$ is a strict interlacer of $\tilde{f}$ satisfying the genericity assumptions (G1)–(G3). Therefore, there exists a sequence $(\tilde{f}_n, \tilde{g}_n)$ converging to (f, g) together with representations $\tilde{f}_n = \det(\widetilde{M}_n)$ with $\tilde{g}_n$ dividing the first principal minor of $\widetilde{M}$ and $\det(\widetilde{M})/\tilde{f}$ a product of $m-d$ linear forms, by Theorem 2.2. The sequence $\widetilde{M}_n$ then has a subsequence converging to a matrix pencil M, which is the desired determinantal representation of f. □

Remark 2.9. The procedure of approximating a given hyperbolic form together with an interlacer as in the proof above may be difficult to carry out in practice. However, the generalized Dixon process can often be applied (with small modifications if needed) even when the genericity assumptions fail.

As a further consequence, we can prove the following rationality result.

Theorem 2.10. *Let $f \in \mathbb{Q}[x, y, z]_d$ be a polynomial hyperbolic with respect to $e \in \mathbb{R}^3$ whose real projective zero set is smooth. Then its hyperbolicity cone is of the form*

$$\{(x, y, z) \in \mathbb{R}^3 : xA + yB + zC \succeq 0\}$$

where A, B, C are symmetric matrices with rational entries of size at most $\binom{d+1}{2}$.

Proof. Let $m \in \mathbb{Q}[x, y, z]^e_{d-1}$ be the vector of all monomials of degree $d-1$, and let $e = \binom{d+1}{2}$. The equation

$$(xA + yB + zC) \cdot m = f \cdot v \tag{2.11}$$

poses linear conditions on the entries of the symmetric $e \times e$ matrices A, B, C and on the entries of $v \in \mathbb{R}^e$. These linear conditions are defined over the rational numbers.

We now apply the above construction to f with a strict interlacer g satisfying the genericity assumptions (G1), (G2), and (G3) whose existence is guaranteed by Lemma 2.1. The vector space V from step (1) is just the vector space of all ternary forms of degree $d-1$, and the a_i form a basis of V. Thus, we can find an invertible matrix $S \in \mathrm{GL}_e(\mathbb{R})$ that maps the vector $a = (a_1, \ldots, a_N)^t$ to the vector m of all monomials of degree $d-1$. In step (7) of our construction we have seen that there is a matrix N such that our symmetric determinantal representation M of $f \cdot h$ satisfies $M \cdot N = \gamma \cdot fh \cdot I_e$. Moreover, the first column of N is $h \cdot a$. Thus, we have that $M \cdot a = \gamma \cdot f \cdot \delta_1$. Now we get the identity

$$S^{-t} M S^{-1} \cdot m = \gamma \cdot f \cdot S^{-t} \delta_1.$$

This is a solution over $\mathbb{R}$ to (2.11) with $e_0 A + e_1 B + e_2 C$ positive definite and $\det(xA + yB + zC) = h \cdot f$ where h is a product of linear forms whose zero set does not intersect the hyperbolicity cone of f.

Since the rational solutions to (2.11) are dense in the solution set over the real numbers, we can find rational matrices A, B, C satisfying (2.11) with $e_0 A + e_1 B + e_2 C$ being positive definite, as well. Then $\det(xA + yB + zC)$ is not the zero polynomial and is divisible by f, since the pencil has a nonzero kernel vector whenever f vanishes at (x, y, z) by (2.11). If A, B, C are chosen close enough to our original solution, the other factor of $\det(xA + yB + zC)$ will not intersect the hyperbolicity cone of f either. □

Remark 2.12. One might be tempted to generalize Theorem 2.10 to singular f using the determinantal representation $\det(M) = h \cdot f$, where h is a product of linear forms, obtained in Corollary 2.8 by a limit argument. In order to make the arguments from the preceding proof work, the first column of M^{adj} would have to be of the form $h \cdot a$ where a is a vector whose entries span a subspace of $\mathbb{R}[x, y, z]_{d-1}$ that is defined over the rationals. It is not clear whether this is always the case.

The next example shows that the smallest size of a rational spectrahedral representation is in general larger than the degree of the curve.

Example 2.13. Consider the univariate polynomial $p = x^3 - 6x - 3 \in \mathbb{Q}[x]$. It has three distinct real zeros but is irreducible over the rational numbers by Eisenstein's criterion. The plane elliptic curve defined by $y^2 = p(x)$ is hyperbolic. Its hyperbolicity cone has the following spectrahedral representation with rational 4×4

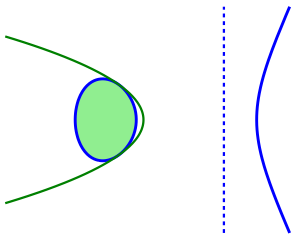


Figure 5. Hyperbolic cubic (in blue), an interlacer touching in 2 real points (in green) and the linear factor (dashed in blue).

matrices:

$$\left\{(x,y,z)\in\mathbb{R}^3 : \begin{pmatrix} 3z & y & -x-z & -3x+z \\ y & -x+2z & 0 & -y \\ -x-z & 0 & z & x+4z \\ -3x+z & -y & x+4z & -x+18z \end{pmatrix} \succeq 0 \right\}.$$

This was obtained by applying our construction to the interlacer $y^2+3xz+z^2$ with two real contact points (Figure 5).

It also has a 3×3 spectrahedral representation with real matrices by the Helton–Vinnikov theorem. It does, however, not have such a representation with rational 3×3 matrices. Indeed, any such representation would yield a contact interlacer defined over the rational numbers by taking some principal 2×2 minor. This interlacer would give rise to a divisor D defined over the rational numbers with $2D = 6P_\infty$ where P_∞ is the point of the curve at infinity. Thus, $D-3P_\infty$ would be an even theta characteristic defined over the rationals. On the other hand, the three even theta characteristics of the curve are given by $P_i - P_\infty$ for P_1, P_2, P_3 the three intersection points of the curve with the x-axis. These are clearly not defined over the rationals.

3. Bézout matrices

Let $f, g \in \mathbb{R}[t]$ be two univariate polynomials having degrees $\deg(f) = d$ and $\deg(g) = d-1$. The *Bézout matrix* of f and g is defined as follows. We write

$$\frac{f(s)g(t)-f(t)g(s)}{s-t} = \sum_{i,j=1}^{d} b_{ij}s^{i-1}t^{j-1}$$

for some real numbers b_{ij}. Then the Bézout matrix is defined as $\mathrm{B}(f,g) = (b_{ij})_{ij}$. Note that $\mathrm{B}(f,g)$ is always a real symmetric matrix. The Bézout matrix can be used to detect the properties of being real-rooted and interlacing.

Theorem 3.1 (see, e.g., [Kreĭn and Naĭmark 1981, §2.2]). *Let $f, g \in \mathbb{R}[t]$ be univariate polynomials with $d = \deg(f) = \deg(g) + 1$. Then the following are equivalent*:

(i) *the Bézout matrix $B(f, g)$ is positive semidefinite and*

(ii) *the polynomial g interlaces f.*

Furthermore, the Bézout matrix has full rank if and only if f and g have no common zero.

In the multivariate case we can proceed analogously. Let $f, g \in \mathbb{R}[x_0, \ldots, x_n]$ be homogeneous polynomials of degrees d and $d-1$, respectively. We assume that f and g do not vanish at $e = (1, 0, \ldots, 0)$. Then, writing $x = (x_1, \ldots, x_n)$, we have

$$\frac{f(s,x)g(t,x) - f(t,x)g(s,x)}{s-t} = \sum_{i,j=1}^{d} b_{ij} s^{i-1} t^{j-1}$$

for some homogeneous polynomials $b_{ij} \in \mathbb{R}[x_1, \ldots, x_n]$ of degree $2d - (i + j)$. Again, we define the Bézout matrix as $\mathrm{B}(f, g) = (b_{ij})_{ij}$. It follows from the above theorem that $\mathrm{B}(f, g)$ is positive definite for every $0 \neq x \in \mathbb{R}^n$ if and only if f is hyperbolic with respect to e and g is a strict interlacer of f.

Remark 3.2. The Bézout matrix $\mathrm{B}(f, g)$ is closely related to the *Wronskian polynomial* $\mathrm{W}(f, g) = \mathrm{D}_e f \cdot g - f \cdot \mathrm{D}_e g$. Namely, if we let $w = (1, x_0, \ldots, x_0^{d-1})^t$, then $\mathrm{W}(f, g) = w^t \cdot \mathrm{B}(f, g) \cdot w$. Indeed, by the definition of the Bézout matrix the right-hand side equals

$$\lim_{s \to t} \left(\frac{f(s,x)g(t,x) - f(t,x)g(s,x)}{s-t} \right) = \mathrm{W}(f, g).$$

We also note that, for square-free polynomials f, the polynomial g of degree $\deg(f) - 1$ is uniquely determined by $\mathrm{W}(f, g)$.

We can use the Wronskian polynomial $\mathrm{W}(f, g)$ to describe the set $\mathrm{Int}(f, e)$ of interlacers of f in direction e, which is a convex cone. By [Kummer et al. 2015, Corollary 2.7], $\mathrm{Int}(f, e)$ can be represented as a linear image of a section of the cone of positive polynomials of degree $2d - 2$, where $d = \deg f$:

$$\mathrm{Int}(f, e) = \{g \in \mathbb{R}[x, y, z]_{d-1} : \mathrm{W}(f, g) \geq 0\}.$$

Whenever $\mathrm{W}(f, g)$ is a sum of squares, the cone $\mathrm{Int}(f, e)$ can be sampled by solving a linear matrix inequality as shown in the following example.

Example 3.3. The cubic $f = x^3 + 2x^2y - xy^2 - 2y^3 - xz^2$ is hyperbolic with respect to $e = (1, 0, 0)$, and $C(f, e)$ is the green region in Figure 6.

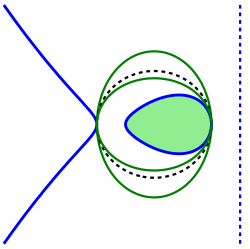


Figure 6. A cubic hyperbolic curve (in blue) with three interlacers, one defined over $\mathbb{Q}$ (in dashed black) and two over an extension of degree 4 (in green). The dashed blue line is the extra factor in the determinantal representation.

Let $g = x^2 + g_{110}xy + g_{101}xz + g_{020}y^2 + g_{011}yz + g_{002}z^2$ be a generic quadratic form such that $g(e) = 1$. The Wronskian of f, g in direction e is the ternary quartic

$$\begin{aligned}\mathrm{W}(f,g) = {} & 2g_{110}x^3y + 2g_{110}x^2y^2 + 2g_{110}y^4 + 2g_{101}x^3z + 2g_{101}x^2yz + 2g_{101}y^3z \\ & + 3g_{020}x^2y^2 + 4g_{020}xy^3 - g_{020}y^4 - g_{020}y^2z^2 + 3g_{011}x^2yz + 4g_{011}xy^2z \\ & - g_{011}y^3z - g_{011}yz^3 + 3g_{002}x^2z^2 + 4g_{002}xyz^2 - g_{002}y^2z^2 - g_{002}z^4 \\ & + x^4 + x^2y^2 + x^2z^2 + 4xy^3.\end{aligned}$$

Let $G = (G_{ij})$ be a symmetric 6×6 matrix of unknowns, and consider the linear system $\mathrm{W}(f,g) = m^t \cdot G \cdot m$, where m is the vector of monomials of degree 2 in x, y, z. We obtain that G (the *Gram matrix* of $\mathrm{W}(f,g)$ [Powers and Wörmann 1998]) has the form

$$G = \begin{bmatrix} 1 & g_{110} & g_{101} & G_{14} & g_{101}-G_{23}+\frac{3}{2}g_{011} & G_{16} \\ g_{110} & H_{22} & G_{23} & 2g_{020}+2 & -G_{34}+2g_{011} & -G_{35}+2g_{002} \\ g_{101} & G_{23} & 1+3g_{002}-2G_{16} & G_{34} & G_{35} & 0 \\ G_{14} & 2g_{020}+2 & G_{34} & 2g_{110}-g_{020} & g_{101}-\frac{1}{2}g_{011} & G_{46} \\ H_{15} & 2g_{011}-G_{34} & G_{35} & g_{101}-\frac{1}{2}g_{011} & -2G_{46}-g_{020}-g_{002} & -\frac{1}{2}g_{011} \\ G_{16} & 2g_{002}-G_{35} & 0 & G_{46} & -\frac{1}{2}g_{011} & -g_{002} \end{bmatrix},$$

where

$$H_{15} = g_{101} - G_{23} + \tfrac{3}{2}g_{011} \quad \text{and} \quad H_{22} = 3g_{020} + 2g_{110} + 1 - 2G_{14}.$$

Let $p_1 = (1,1,0)$ and $p_2 = (1,-1,0)$. Interlacers in $\mathrm{Int}(f,e)$ vanishing in p_1 and p_2 can be computed through the quantified linear matrix inequality

$$\text{there exists } G_{ij} : g(p_1) = g(p_2) = 0,\ G \succeq 0. \tag{3.4}$$

Solving (3.4) symbolically using [Henrion et al. 2019] yields the parametrization of an interlacer $g = x^2 - y^2 + t \cdot z^2$, where t is any of the two real roots t_1, t_2 of $q(t) = 49t^4 - 20t^3 + 22t^2 + 12t + 1$ (the green curves in Figure 6).

Since the matrices G corresponding to the two interlacers have rank two, the corresponding Wronskian polynomials are sums of two squares. Choosing a rational $t_1 < r < t_2$ gives a rational interlacer, for instance $g = x^2 - y^2 - \frac{1}{5}z^2$.

As in Example 2.13, our construction yields rational 4×4 determinantal representations of f times a rational linear polynomial that can be built from the interlacer $g = x^2 - y^2 - \frac{1}{5}z^2$:

$$\tfrac{24}{125} f \cdot (2x - y) = \det \begin{pmatrix} 5x + 10y & -x - 2y & -4z & 2z \\ -x - 2y & x & 0 & 0 \\ -4z & 0 & 4x + 2y & -2x - 4y \\ 2z & 0 & -2x - 4y & 4x + 2y \end{pmatrix}.$$

The matrix on the right-hand side of the previous equality gives a spectrahedral representation of $C(f, e)$ (the green region in Figure 6).

In the following, we show how our construction gives a *sum-of-squares decomposition*, i.e., a representation $B(f, g) = S^t S$ for some (not necessarily square) matrix S with polynomial entries, for any curve f hyperbolic with respect to $(1, 0, 0)$ and any strict interlacer g.

We have seen that there is a basis $g_1, \ldots, g_N$ of $\mathbb{R}[x, y, z]_{d-1}$ with $g_1 = g$ and real symmetric matrices A, B, C of size N such that A is positive definite and

$$(xA + yB + zC) \cdot v = \delta_1 \cdot f \tag{3.5}$$

where $v = (g_1, \ldots, g_N)^t$ and $\delta_1 \in \mathbb{R}^N$ is the first unit vector. Let us write $v = h_0 x^{d-1} + \cdots + h_{d-1}$ for some $h_i \in \mathbb{R}[y, z]_i^N$, and let S be the matrix with columns $h_d, \ldots, h_0$. We claim that $\mathrm{B}(f, g) = S^t A S$. Indeed, by [Kummer 2017, §3], we have that $\mathrm{B}(f, \tilde{g}) = S^t A S$ for some $\tilde{g} \in \mathbb{R}[x, y, z]_{d-1}$. Furthermore, taking the derivative of (3.5) yields

$$A \cdot v + (xA + yB + zC) \cdot \mathrm{D}_e v = \delta_1 \cdot \mathrm{D}_e f.$$

Now it follows by multiplying with v^t from the left and another application of (3.5) that

$$v^t \cdot A \cdot v + f \cdot \delta_1^t \cdot \mathrm{D}_e v = v^t \cdot A \cdot v + v^t \cdot (xA + yB + zC) \cdot \mathrm{D}_e v = v^t \cdot \delta_1 \cdot \mathrm{D}_e f.$$

Thus, by Remark 3.2, applied with $v = Sw$, we find

$$\mathrm{W}(f, \tilde{g}) = w^t \mathrm{B}(f, \tilde{g}) w = w^t S^t A S w = v^t \cdot \delta_1 \mathrm{D}_e f - f \cdot \delta_1^t \mathrm{D}_e v = \mathrm{W}(f, g),$$

which implies $g = \tilde{g}$, again by Remark 3.2, since f is square-free.

Remark 3.6. It has been shown in [Kummer 2017] that a sum-of-squares representation $\mathrm{B}(f, g) = S^t A S$ of a Bézout matrix of a hyperbolic polynomial f with a positive definite matrix A as above gives rise to a definite determinantal representation of some multiple of f. Now we have seen that for every strict interlacer of a hyperbolic curve there is a sum-of-squares decomposition of the corresponding Bézout matrix which even gives rise to a spectrahedral representation of the hyperbolicity cone.

Acknowledgements

We thank the Max Planck Institute for Mathematics in the Sciences in Leipzig for its hospitality, where this work was started. We would also like to thank the referee for useful remarks that helped us improve the exposition. Simone Naldi acknowledges the support of the Fondation Mathématique Jacques Hadamard through the PGMO project 2018-0061H.

References

[Arbarello et al. 1985] E. Arbarello, M. Cornalba, P. A. Griffiths, and J. Harris, *Geometry of algebraic curves*, vol. I, Grundlehren der Mathematischen Wissenschaften **267**, Springer, 1985. MR Zbl

[Beauville 2000] A. Beauville, "Determinantal hypersurfaces", *Michigan Math. J.* **48** (2000), 39–64. MR Zbl

[Dixon 1902] A. C. Dixon, "Note on the reduction of a ternary quantic to a symmetrical determinant", *P. Cambridge Philos. Soc.* **11** (1902), 350–351. Zbl

[Dolgachev 2012] I. V. Dolgachev, *Classical algebraic geometry: a modern view*, Cambridge University, 2012. MR Zbl

[Fulton 1989] W. Fulton, *Algebraic curves: an introduction to algebraic geometry*, Addison-Wesley, Redwood City, CA, 1989. MR Zbl

[Helton and Vinnikov 2007] J. W. Helton and V. Vinnikov, "Linear matrix inequality representation of sets", *Comm. Pure Appl. Math.* **60**:5 (2007), 654–674. MR Zbl

[Henrion et al. 2019] D. Henrion, S. Naldi, and M. Safey El Din, "SPECTRA — a Maple library for solving linear matrix inequalities in exact arithmetic", *Optim. Methods Softw.* **34**:1 (2019), 62–78. MR Zbl

[Kreĭn and Naĭmark 1981] M. G. Kreĭn and M. A. Naĭmark, "The method of symmetric and Hermitian forms in the theory of the separation of the roots of algebraic equations", *Linear and Multilinear Algebra* **10**:4 (1981), 265–308. MR Zbl

[Kummer 2017] M. Kummer, "Determinantal representations and Bézoutians", *Math. Z.* **285**:1–2 (2017), 445–459. MR Zbl

[Kummer et al. 2015] M. Kummer, D. Plaumann, and C. Vinzant, "Hyperbolic polynomials, interlacers, and sums of squares", *Math. Program. B* **153**:1 (2015), 223–245. MR Zbl

[Plaumann and Vinzant 2013] D. Plaumann and C. Vinzant, "Determinantal representations of hyperbolic plane curves: an elementary approach", *J. Symbolic Comput.* **57** (2013), 48–60. MR Zbl

[Powers and Wörmann 1998] V. Powers and T. Wörmann, "An algorithm for sums of squares of real polynomials", *J. Pure Appl. Algebra* **127**:1 (1998), 99–104. MR Zbl

[Vinnikov 2012] V. Vinnikov, "LMI representations of convex semialgebraic sets and determinantal representations of algebraic hypersurfaces: past, present, and future", pp. 325–349 in *Mathematical methods in systems, optimization, and control*, edited by H. Dym et al., Operator Theory: Advances and Applications **222**, Springer, 2012. MR Zbl

Received August 28, 2018. Revised May 30, 2019.

MARIO KUMMER
INSTITUT FÜR MATHEMATIK
TECHNISCHE UNIVERSITÄT BERLIN
BERLIN
GERMANY
kummer@tu-berlin.de

SIMONE NALDI
XLIM
UNIVERSITÉ DE LIMOGES
LIMOGES
FRANCE
simone.naldi@unilim.fr

DANIEL PLAUMANN
FAKULTÄT FÜR MATHEMATIK
TECHNISCHE UNIVERSITÄT DORTMUND
DORTMUND
GERMANY
daniel.plaumann@math.tu-dortmund.de

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.265

DEFORMATIONS OF LINEAR LIE BRACKETS

PIER PAOLO LA PASTINA AND LUCA VITAGLIANO

A VB-algebroid is a vector bundle object in the category of Lie algebroids. We attach to every VB-algebroid a differential graded Lie algebra and we show that it controls deformations of the VB-algebroid structure. Several examples and applications are discussed. This is the first in a series of papers devoted to deformations of vector bundles and related structures over differentiable stacks.

Introduction

Lie algebroids are ubiquitous in differential geometry: they encompass several algebraic and geometric structures such as Lie algebras, tangent bundles, foliations, Poisson brackets, Lie algebra actions on manifolds and so on, and they are the infinitesimal counterparts of Lie groupoids. The notion of Lie algebroid appeared for the first time in the work of Pradines [1967] and has become more and more important in the last fifty years. In particular, deformations of Lie algebroids have been discussed by Crainic and Moerdijk [2008], while deformations of Lie groupoids have been studied very recently by Crainic, Mestre and Struchiner [Crainic et al. 2015].

VB-algebroids are vector bundle objects in the category of Lie algebroids [Mackenzie 1998a; Gracia-Saz and Mehta 2010]. They emerge naturally in the study of Lie algebroids. For instance, the tangent and the cotangent bundles of a Lie algebroid are VB-algebroids. Additionally, VB-algebroids are generalizations of ordinary representations of Lie algebroids: specifically they are equivalent to 2-term representations up to homotopy of Lie algebroids, hence to (special kinds of) representations of Lie algebroids on graded vector bundles [Arias Abad and Crainic 2012; Gracia-Saz and Mehta 2010]. Finally, VB-algebroids are the infinitesimal counterparts of VB-groupoids. The latter serve as models for vector bundles over certain singular spaces: *differentiable stacks* [Behrend and Xu 2011]. Examples of differentiable stacks are orbifolds, leaf spaces of foliations and orbit spaces of Lie group actions.

This is the first in a series of papers devoted to deformations of vector bundles over differentiable stacks and related deformation problems. A first step in this

MSC2010: primary 22A22, 58H05, 58H15; secondary 17B55.
Keywords: Lie algebroids, VB-algebroids, deformations, graded manifolds, deformation cohomology.

direction has been taken by del Hoyo and Ortiz [2016], who have shown that the VB-cohomology of a VB-groupoid is actually *VB-Morita invariant*, i.e., it is an invariant of the associated vector bundle of differentiable stacks. Notice that several important geometric structures, like Riemannian metrics, symplectic forms, complex structures, etc., can be seen as vector bundle maps. In order to study deformations of the former, it is then useful to study deformations of vector bundles themselves first. In this paper, we begin this program working at the infinitesimal level, i.e., studying deformations of VB-algebroids. More precisely, we study deformations of VB-algebroid structures on double vector bundles. In the second paper of the series we will study deformations of VB-groupoids and their behavior under the Lie functor [La Pastina and Vitagliano 2019].

The paper is divided into two main sections. The first one presents the general theory, and the second one discusses examples and applications. In turn, the first section is divided into four subsections. In Section 1.1 we recall from [Crainic and Moerdijk 2008] the differential graded Lie algebra (DGLA) controlling deformations of Lie algebroids. We also discuss gauge equivalent deformations, something that is missing in the original discussion by Crainic and Moerdijk. In Section 1.2 we recall the basics of VB-algebroids, in particular their description in terms of graded manifolds. In Section 1.3 we discuss deformations of VB-algebroids. Let $(W \Rightarrow E; A \Rightarrow M)$ be a VB-algebroid. In particular $W \Rightarrow E$ is a Lie algebroid, so it has an associated deformation DGLA. We show that deformations of $(W \Rightarrow E; A \Rightarrow M)$ are controlled by the sub-DGLA of *linear* cochains, originally introduced in [Esposito et al. 2016], that we call the *linear deformation complex*, and we provide various equivalent descriptions of this object. The most efficient one involves the *homogeneity structure* of $W \to A$, i.e., the action of the monoid $\mathbb{R}_{\geq 0}$ on the total space by fiber-wise homotheties: linear deformation cochains are precisely those that are invariant under the action by (nonzero) homotheties. It is clear that this action induces graded subalgebras of the algebras of functions, differential forms and multivectors on the total space of a vector bundle and can be used to define *linear* objects in these algebras, thus giving a unified framework to the original definitions in [Bursztyn and Cabrera 2012] and [Iglesias-Ponte et al. 2012]. We recall this briefly in the Appendix. Another important description of the linear deformation complex is in terms of graded geometry. It is well known that Lie algebroids are equivalent to DG-manifolds concentrated in degree 0 and 1 and VB-algebroids are equivalent to vector bundles in the category of such graded manifolds [Mehta 2006; Vaintrob 1997; Voronov 2012]. Moreover, it is (implicitly) shown in [Crainic and Moerdijk 2008] that the deformation DGLA of a Lie algebroid $A \Rightarrow M$ is isomorphic to the DGLA of vector fields on $A[1]$, giving an elegant and manageable interpretation. A similar interpretation becomes very useful in the case of VB-algebroids.

In Section 1.4 we show that it is possible to "linearize" deformation cochains of the top algebroid $W \Rightarrow E$ of a VB-algebroid $(W \Rightarrow E; A \Rightarrow M)$, adapting a technique from [Cabrera and Drummond 2017]. The main consequence is that the linear deformation cohomology is embedded, as a graded Lie algebra, in the deformation cohomology of the top algebroid.

In the second section of the paper we present examples. We discuss in detail particularly simple instances of VB-algebroids coming from linear algebra, namely VB-algebras and LA-vector spaces (Sections 2.1 and 2.2 respectively). VB-algebras are equivalent to Lie algebra representations, and our discussion encompasses the classical theory of Nijenhuis and Richardson [1966; 1967a]. In Section 2.3, we discuss deformations of the tangent and the cotangent VB-algebroids of a Lie algebroid. Partial connections along foliations and Lie algebra actions on vector bundles can be also encoded by VB-algebroids and we study the associated deformation complexes in Sections 2.4 and 2.5 respectively. We also discuss VB-algebroids of type 1 in the classification of Gracia-Saz and Mehta [2010]. Their deformation cohomology is canonically isomorphic to that of the base algebroid (Section 2.6).

We usually indicate with a bullet $^\bullet$ the presence of a degree in a graded vector space. If $V^\bullet$ is a graded vector space, its *shift by one* $V[1]^\bullet$ is defined by $V[1]^k = V^{k+1}$. We assume the reader is familiar with graded manifolds and the graded geometry description of Lie algebroids. Here, we only recall that a graded manifold is *concentrated* in degree $k, \dots, k+l$, if the degrees of its coordinates range from k to $k+l$ and a DG-manifold is a graded manifold equipped with an homological vector field. For instance, if $A \Rightarrow M$ is a Lie algebroid, then shifting by one the degree of the fibers of the vector bundle $A \to M$, we get a DG-manifold $A[1]$, concentrated in degree 0 and 1, whose homological vector field is the de Rham differential d_A of A. Explicitly, the algebra of smooth functions on $A[1]$ is

$$C^\infty(A[1]) = \Omega^\bullet_A := \Gamma(\wedge^\bullet A^*).$$

The correspondence $A \rightsquigarrow A[1]$ establishes an equivalence between the category of Lie algebroids and the category of DG-manifolds concentrated in degree 0 and 1 [Vaintrob 1997]. We stress that the graded manifold $A[1]$ is obtained from A by *assigning a degree* 1 to the linear fiber coordinates. We warn the unfamiliar reader that, despite the notation, the shift $A \rightsquigarrow A[1]$ is (related but) different from the degree shift for a graded vector space discussed at the beginning of this paragraph. The reader can find more details in [Mehta 2006] which is also our main reference for graded geometry.

1. Deformations of VB-algebroids

1.1. ***Deformations of Lie algebroids.*** A Lie algebroid $A \Rightarrow M$ over a manifold M is a vector bundle $A \to M$ with a Lie bracket $[-,-]$ on its space of sections $\Gamma(A)$

and a bundle map $\rho : A \to TM$, satisfying the Leibniz rule:

$$[\alpha, f\beta] = \rho(\alpha)(f)\beta + f[\alpha, \beta]$$

for all $\alpha, \beta \in \Gamma(A)$, $f \in C^\infty(M)$.

We briefly recall the deformation theory of Lie algebroids, originally due to Crainic and Moerdijk [2008], adding some small details about equivalence of deformations which are missing in the original treatment. We begin with a vector bundle $E \to M$. Let $k \geq 0$.

Definition 1.1.1. A *multiderivation* of E with k entries (and $C^\infty(M)$-multilinear symbol), also called a *k-derivation*, is a skew-symmetric, $\mathbb{R}$-k-linear map

$$c : \Gamma(E) \times \cdots \times \Gamma(E) \to \Gamma(E)$$

such that there exists a bundle map $\sigma_c : \wedge^{k-1} E \to TM$, the *symbol* of c, satisfying the following Leibniz rule:

$$c(\alpha_1, \ldots, \alpha_{k-1}, f\alpha_k) = \sigma_c(\alpha_1, \ldots, \alpha_{k-1})(f)\alpha_k + fc(\alpha_1, \ldots, \alpha_k),$$

for all $\alpha_1, \ldots, \alpha_k \in \Gamma(E)$, $f \in C^\infty(M)$.

1-derivations are simply *derivations*, 2-derivations are called *biderivations*. The space of derivations of E is denoted by $\mathfrak{D}(E)$ (or $\mathfrak{D}(E, M)$ if we want to insist on the fact that the base of the vector bundle E is M). The space of k-derivations is denoted $\mathfrak{D}^k(E)$ (or $\mathfrak{D}^k(E, M)$). In particular, $\mathfrak{D}^1(E) = \mathfrak{D}(E)$. We also put $\mathfrak{D}^0(E) = \Gamma(E)$ and $\mathfrak{D}^\bullet(E) = \bigoplus_{k\geq 0} \mathfrak{D}^k(E)$. Then $\mathfrak{D}^\bullet(E)[1]$, endowed with the *Gerstenhaber bracket* $[\![-, -]\!]$, is a graded Lie algebra. We recall that, for $c_1 \in \mathfrak{D}^k(E)$, and $c_2 \in \mathfrak{D}^l(E)$, the *Gerstenhaber product* of c_1 and c_2 is the $\mathbb{R}$-$(k + l - 1)$-linear map $c_1 \circ c_2$ given by

$$\begin{aligned}(c_1 \circ c_2)(\alpha_1, \ldots, \alpha_{k+l-1}) \\ = \sum_{\tau \in S_{l,k-1}} (-1)^\tau c_1(c_2(\alpha_{\tau(1)}, \ldots, \alpha_{\tau(l)}), \alpha_{\tau(l+1)}, \ldots, \alpha_{\tau(l+k-1)})\end{aligned}$$

for all $\alpha_1, \ldots, \alpha_{k+l-1} \in \Gamma(E)$, and the Gerstenhaber bracket is defined by

$$[\![c_1, c_2]\!] = (-1)^{(k-1)(l-1)} c_1 \circ c_2 - c_2 \circ c_1.$$

The graded Lie algebra $\mathfrak{D}^\bullet(E)[1]$ first appeared in [Grabowska et al. 2003].

The group of vector bundle automorphisms of E acts naturally on multiderivations of E. If $\phi : E \to E$ is an automorphism covering the diffeomorphism $\phi_M : M \to M$, then ϕ acts on sections of E (by pull-back) via the following formula:

$$\phi^*\alpha := \phi^{-1} \circ \alpha \circ \phi_M, \quad \alpha \in \Gamma(E),$$

and it acts on higher degree multiderivations via:

$$(\phi^* c)(\alpha_1, \ldots, \alpha_k) := \phi^*(c(\phi^{-1*}\alpha_1, \ldots, \phi^{-1*}\alpha_k))$$

for all $\alpha_1, \ldots, \alpha_k \in \Gamma(E)$, $c \in \mathfrak{D}^k(E)$. Moreover, ϕ acts in the obvious way on sections of the dual bundle E^*. It is clear that

$$\begin{aligned} \phi^*(f\alpha) &= \phi_M^* f \cdot \phi^*\alpha, \\ \phi^*(c(\alpha_1, \ldots, \alpha_k)) &= (\phi^* c)(\phi^*\alpha_1, \ldots, \phi^*\alpha_k), \\ \phi_M^*\langle \varphi, \alpha\rangle &= \langle \phi^*\varphi, \phi^*\alpha\rangle, \end{aligned} \tag{1-1}$$

for all $\alpha, \alpha_1, \ldots, \alpha_k \in \Gamma(E)$, $f \in C^\infty(M)$, and $\varphi \in \Gamma(E^*)$, where $\langle -, -\rangle : E^* \otimes E \to \mathbb{R}$ is the duality pairing. Finally, ϕ acts on the exterior algebras of E and E^*, and it also acts on vector bundle maps $\wedge^\bullet E \to TM$ in the obvious way.

A direct computation shows that the action of vector bundle automorphisms on multiderivations does also respect the Gerstenhaber bracket, i.e.,

$$\phi^*[\![c_1, c_2]\!] = [\![\phi^* c_1, \phi^* c_2]\!] \tag{1-2}$$

for all $c_1, c_2 \in \mathfrak{D}^\bullet(E)$. Additionally,

$$\phi^*\sigma_c = \sigma_{\phi^* c} \tag{1-3}$$

for all $c \in \mathfrak{D}^\bullet(E)$.

If $A \Rightarrow M$ is a Lie algebroid, the Lie bracket $b_A = [-, -]$ on sections of A is a biderivation and it contains the full information about $A \Rightarrow M$. Additionally, $[\![b_A, b_A]\!] = 0$ as a consequence of the Jacobi identity. We summarize this remark with the following:

Proposition 1.1.2. *Lie algebroid structures on $A \to M$ are in one-to-one correspondence with Maurer–Cartan elements in the graded Lie algebra $\mathfrak{D}^\bullet(A)[1]$, i.e., degree 1 elements b such that $[\![b, b]\!] = 0$.*

Now, fix a Lie algebroid structure $A \Rightarrow M$ on the vector bundle $A \to M$, and let b_A be the Lie bracket on sections of A. Equipped with the Gerstenhaber bracket and the interior derivation $\delta := [\![b_A, -]\!]$, $\mathfrak{D}^\bullet(A)[1]$ is a differential graded Lie algebra (DGLA), denoted $C^\bullet_{\mathrm{def}}(A)$ (or $C^\bullet_{\mathrm{def}}(A, M)$ if we want to insist on the base manifold being M) and called the *deformation complex of A*. The cohomology of $C^\bullet_{\mathrm{def}}(A)$ is denoted $H^\bullet_{\mathrm{def}}(A)$ (or $H^\bullet_{\mathrm{def}}(A, M)$), and called the *deformation cohomology* of A.

Remark 1.1.3. Notice that we adopt a different convention from that of [Crainic and Moerdijk 2008], where $C^k_{\mathrm{def}}(A)$ is the space of k-derivations. With that convention, however, $C^\bullet_{\mathrm{def}}(A)$ is a DGLA only up to a shift.

The differential $\delta : C^\bullet_{\mathrm{def}}(A) \to C^{\bullet+1}_{\mathrm{def}}(A)$ is given, on k-derivations, by

$$\begin{aligned}(1\text{-}4)\quad \delta c(\alpha_1,\dots,\alpha_{k+1}) = &\sum_i (-1)^{i+1}[\alpha_i, c(\alpha_1,\dots,\widehat{\alpha_i},\dots,\alpha_{k+1})] \\ &+\sum_{i<j}(-1)^{i+j} c([\alpha_i,\alpha_j],\alpha_1,\dots,\widehat{\alpha_i},\dots,\widehat{\alpha_j},\dots,\alpha_{k+1}).\end{aligned}$$

Definition 1.1.4. A *deformation* of b_A is any (other) Lie algebroid structure on the vector bundle $A \to M$.

It is clear that $b = b_A + c$ satisfies $[\![b, b]\!] = 0$ if and only if

$$\delta c + \tfrac{1}{2}[\![c, c]\!] = 0,$$

i.e., c is a (degree 1) solution of the *Maurer–Cartan equation* in the DGLA $C^\bullet_{\mathrm{def}}(A)$. Hence Proposition 1.1.2 can be rephrased saying that *deformations of b_A are in one-to-one correspondence with Maurer–Cartan elements of $C^\bullet_{\mathrm{def}}(A)$.*

Now, let b_0, b_1 be deformations of b_A. We say that b_0 and b_1 are *equivalent* if there exists a *fiber-wise linear* isotopy taking b_0 to b_1, i.e., there is a smooth path of vector bundle automorphisms $\phi_t : A \to A,\ t \in [0, 1]$, such that $\phi_0 = \mathrm{id}_A$ and $\phi_1^* b_1 = b_0$. On the other hand, two Maurer–Cartan elements c_0, c_1 are *gauge-equivalent* if they are interpolated by a smooth path of 1-cochains c_t, and c_t is a solution of the following ODE:

$$\frac{dc_t}{dt} = \delta\Delta_t + [\![c_t, \Delta_t]\!] \tag{1-5}$$

for some smooth path of 0-cochains (i.e., derivations) $\Delta_t,\ t \in [0, 1]$.

Notice that (1-5) is equivalent to

$$\frac{db_t}{dt} = [\![b_t, \Delta_t]\!], \tag{1-6}$$

where $b_t = b_A + c_t$.

Proposition 1.1.5. *Let $b_0 = b_A + c_0$, $b_1 = b_A + c_1$ be deformations of b_A. If b_0, b_1 are equivalent, then c_0, c_1 are gauge-equivalent. If M is compact, the converse is also true.*

Proof. Suppose that b_0 and b_1 are equivalent deformations, and let $\phi_t : A \to A$ be a fiber-wise linear isotopy taking b_0 to b_1. Set $b_t = \phi_t^{-1*} b_0 = b_A + c_t$, and let Δ_t be the infinitesimal generator of ϕ_t, i.e.,

$$\frac{d\phi_t^*}{dt} = \phi_t^* \circ \Delta_t. \tag{1-7}$$

Notice that

$$[\![b_t, b_t]\!] = [\![\phi_t^{-1*} b_0, \phi_t^{-1*} b_0]\!] = \phi_t^{-1*}[\![b_0, b_0]\!] = 0,$$

so b_t is a deformation of b_A for all t. Moreover, $\phi_t^*(b_t(\alpha,\beta)) = b_0(\phi_t^*\alpha, \phi_t^*\beta)$ for all $\alpha, \beta \in \Gamma(A)$. Differentiating with respect to t, we obtain:

$$\begin{aligned}\phi_t^*\left(\Delta_t(b_t(\alpha,\beta)) + \frac{db_t}{dt}(\alpha,\beta)\right) &= b_0(\phi_t^*(\Delta_t(\alpha)), \phi_t^*\beta) + b_0(\phi_t^*\alpha, \phi_t^*(\Delta_t(\beta))) \\ &= \phi_t^*(b_t(\Delta_t(\alpha),\beta) + b_t(\alpha, \Delta_t(\beta))),\end{aligned}$$

so

$$\frac{db_t}{dt}(\alpha,\beta) = b_t(\Delta_t(\alpha),\beta) + b_t(\alpha,\Delta_t(\beta)) - \Delta_t(b_t(\alpha,\beta)),$$

i.e., (1-6), hence (1-5), holds, as desired.

Conversely, suppose that M is compact and there exist a family of derivations Δ_t and a family of 1-cochains b_t such that (1-5) or, equivalently, (1-6) holds. Let X_t be the symbol of Δ_t. From compactness, X_t is a complete time-dependent vector field on M, i.e., it generates a complete flow $(\phi_M)_t$. The time dependent derivation Δ_t generates a flow by vector bundle automorphisms $\phi_t : A \to A$, covering the complete flow $(\phi_M)_t$ (and implicitly defined by the ODE (1-7)). By linearity, ϕ_t is a complete flow itself. We want to show that

$$\phi_t^*(b_t(\alpha,\beta)) = b_0(\phi_t^*\alpha, \phi_t^*\beta), \quad \alpha,\beta \in \Gamma(A). \tag{1-8}$$

For $t = 0$ this is obviously true and the derivatives of both sides are the same because of (1-6). So we have (1-8), and, by taking $t = 1$, we conclude that ϕ_t is a (fiber-wise linear) isotopy taking b_0 to b_1. □

Remark 1.1.6. An *infinitesimal deformation* of a Lie algebroid $A \Rightarrow M$ is an element $c \in C^1_{\mathrm{def}}(A)$ such that $\delta c = 0$, i.e., a 1-cocycle in $C^\bullet_{\mathrm{def}}(A)$. As usual in deformation theory, this definition is motivated by the fact that, *if c_t is a smooth path of Maurer–Cartan elements starting at 0, then $(dc_t/dt)|_{t=0}$ is an infinitesimal deformation of A*. More generally, the cocycle condition $\delta c = 0$ is just the linearization at $c = 0$ of the Maurer–Cartan equation. Hence, 1-cocycles in $C^\bullet_{\mathrm{def}}(A)$ can be seen as the (formal) tangent vectors to the variety of Maurer–Cartan elements. Similarly, 1-coboundaries can be seen as tangent vectors to the gauge orbit through 0. We conclude that $H^1_{\mathrm{def}}(A)$ is the *formal tangent space* to the moduli space of deformations under gauge equivalence.

Remark 1.1.7. The deformation complex of a Lie algebroid has an efficient description in terms of graded geometry. In fact, graded geometry becomes very useful when dealing with several issues related to VB-algebroids.

Let $A \Rightarrow M$ be a Lie algebroid and let $(\Omega^\bullet_A = \Gamma(\wedge^\bullet A^*), d_A)$ be its de Rham complex (sometimes we will use $\Omega^\bullet_{A,M}$ for A-forms, if we want to insist on M being the base manifold). Cochains in $\Omega^\bullet_A$ can be seen as functions on the DG-manifold $A[1]$ obtained from A shifting by one the fiber degree. The Q-structure on $A[1]$ is simply d_A. Additionally, there is a canonical isomorphism $C^\bullet_{\mathrm{def}}(A) \cong \mathfrak{X}(A[1])^\bullet$ of

DGLAs, where $\mathfrak{X}(A[1])^\bullet$ is the space of vector fields on the DG-manifold $A[1]$ (in other words, $\mathfrak{X}(A[1])^\bullet$ is the space of (graded) derivations of $\Omega^\bullet_A$). With the graded commutator and the adjoint operator $[d_A, -]$, $\mathfrak{X}(A[1])^\bullet$ is indeed a DGLA. The isomorphism $C^\bullet_{\mathrm{def}}(A) \to \mathfrak{X}(A[1])^\bullet$, $c \mapsto \delta_c$ can be described explicitly as follows. Let $c \in C^k(A)$ and let σ_c be the symbol of c. Then $\delta_c \in \mathfrak{X}(A[1])^\bullet$ is the degree k vector field that takes $\omega \in \Omega^p_A$, to $\delta_c\omega \in \Omega^{k+p}_A$ with

$$\begin{aligned}(1\text{-}9)\quad \delta_c\omega(\alpha_1,\dots,\alpha_{k+p}) = &\sum_{\tau\in S_{k,p}} (-1)^\tau \sigma_c(\alpha_{\tau(1)},\dots,\alpha_{\tau(k)})\omega(\alpha_{\tau(k+1)},\dots,\alpha_{\tau(k+p)})\\ &- \sum_{\tau\in S_{k+1,p-1}} (-1)^\tau \omega(c(\alpha_{\tau(1)},\dots,\alpha_{\tau(k+1)}),\alpha_{\tau(k+2)},\dots,\alpha_{\tau(k+p)}),\end{aligned}$$

where $S_{l,m}$ denotes (l,m)-unshuffles. Notice that c can be reconstructed from δ_c by using formula (1-9) for $p = 0, 1$:

$$\delta_c f(\alpha_1,\dots,\alpha_k) = \sigma_c(\alpha_1,\dots,\alpha_k) f, \tag{1-10}$$

and

$$\begin{aligned}(1\text{-}11)\quad &\delta_c\varphi(\alpha_1,\dots,\alpha_{k+1}) =\\ &\sum_i (-1)^{k-i}\sigma_c(\alpha_1,\dots,\widehat{\alpha_i},\dots,\alpha_{k+1})\langle\varphi,\alpha_i\rangle + \langle\varphi, c(\alpha_1,\dots,\alpha_{k+1})\rangle,\end{aligned}$$

where $f \in C^\infty(M)$, $\varphi \in \Omega^1_A = \Gamma(A^*)$, and $\alpha_1,\dots,\alpha_{k+1} \in \Gamma(A)$.

1.2. ***Double vector bundles and VB-algebroids.*** In this section we recall the basic definitions and properties of double vector bundles and VB-algebroids that will be useful later. For all the necessary details about the homogeneity structure of a vector bundle, including our notations, we refer to the Appendix, which we recommend reading before continuing with the bulk of the paper. We only recall here that, given a vector bundle $E \to M$, the *homogeneity structure* of E is the action $h : \mathbb{R}_{\geq 0} \times E \to E$, $(\lambda, e) \mapsto h_\lambda e := \lambda \cdot e$, of nonnegative reals on E by homotheties (fiber-wise multiplication by scalars).

Definition 1.2.1. A *double vector bundle* (DVB for short) is a vector bundle in the category of vector bundles. More precisely, it is a commutative square

$$\begin{array}{ccc} W & \xrightarrow{\tilde{p}} & E \\ {\scriptstyle q_W}\downarrow & & \downarrow{\scriptstyle q} \\ A & \xrightarrow{p} & M \end{array} \tag{1-12}$$

where all four sides are vector bundles, the projection $q_W : W \to A$, the addition $+_A : W \times_A W \to W$, the multiplication $\lambda \cdot_A : W \to W$ by any scalar $\lambda \in \mathbb{R}$ in the fibers of $W \to A$ and the zero section $\tilde{0}^A : A \to W$ are vector bundle maps covering

the projection $q : E \to M$, the addition $+ : E \times_M E \to E$, the scalar multiplication $\lambda \cdot : E \to E$ and the zero section $0^E : M \to E$, respectively. The projection, the addition, the scalar multiplication and the zero section of a vector bundle will be called the *structure maps*. DVB (1-12) will be also denoted by $(W \to E;\, A \to M)$.

Notice that W is a vector bundle over E and over A, so it carries two homogeneity structures. However, we will mainly use the latter and denote it simply by h. For many more details on DVBs we refer to [Mackenzie 2005] and [Gracia-Saz and Mehta 2010].

Let $(W \to E;\, A \to M)$ be a DVB. The manifold W will be called the *total space*. Consider the submanifold

$$C := \ker(W \to E) \cap \ker(W \to A) \subset W.$$

In other words, elements of C are those projecting simultaneously on the (images of the) zero sections of A and E (which are both diffeomorphic to M). The fiberwise operations of the vector bundles $W \to E$ and $W \to A$ coincide on C (see [Mackenzie 2005]), so they define a (unique) vector bundle structure on C over M. The vector bundle $C \to M$ is called the *core* of $(W \to E;\, A \to M)$.

In the following, we denote by $\Gamma(W, E)$ the space of sections of $W \to E$. Sections of $C \to M$ can be naturally embedded into $\Gamma(W, E)$, via the map $\Gamma(C) \to \Gamma(W, E)$, $\chi \mapsto \bar{\chi}$, defined by:

$$\bar{\chi}_e = \tilde{0}^E_e +_A \chi_{q(e)}, \quad e \in E. \tag{1-13}$$

The image of the inclusion $\chi \mapsto \bar{\chi}$ is, by definition, the space $\Gamma_{\text{core}}(W, E)$ of *core sections* of $W \to E$.

There is another relevant class of sections of $W \to E$: *linear sections*. We say that a section of $W \to E$ is a *linear section* if it is a vector bundle map covering some section of $A \to M$. The space of linear sections of $W \to E$ is denoted $\Gamma_{\text{lin}}(W, E)$. We will usually denote by $\tilde{\alpha}, \tilde{\beta}, \ldots$ the sections in $\Gamma_{\text{lin}}(W, E)$. The $C^\infty(E)$-module $\Gamma(W, E)$ is spanned by $\Gamma_{\text{core}}(W, E)$ and $\Gamma_{\text{lin}}(W, E)$.

Linear and core sections of $W \to E$ can be efficiently characterized using the homogeneity structure h. Namely, the following lemma holds.

Lemma 1.2.2. *A section $w \in \Gamma(W, E)$ is*

(1) *linear if and only if $h_\lambda^* w = w$ for every $\lambda > 0$;*

(2) *core if and only if $h_\lambda^* w = \lambda^{-1} w$ for every $\lambda > 0$.*

More generally, we say that a section w of $W \to E$ is *of weight q* if $h_\lambda^* w = \lambda^q w$ for every $\lambda > 0$. Using this terminology, linear sections are precisely sections of weight 0 and core sections are sections of weight -1. It is easy to check that *there are no nonzero sections of $W \to E$ of weight less than -1.*

Remark 1.2.3. Let $(W \to E;\, A \to M)$ be a DVB, let C be its core and let $W_A^* \to A$ be the dual vector bundle of $W \to A$. Then

$$\begin{array}{ccc} W_A^* & \longrightarrow & C^* \\ \downarrow & & \downarrow \\ A & \longrightarrow & M \end{array}$$

is a DVB, called the *dual* of W over A, whose core is E^*. We refer to [Mackenzie 2005] for the structure maps of the dual DVB.

Example 1.2.4. A distinguished example of a DVB is the *tangent double of a vector bundle*. If $E \to M$ is a vector bundle, then

$$\begin{array}{ccc} TE & \longrightarrow & E \\ \downarrow & & \downarrow \\ TM & \longrightarrow & M \end{array}$$

is a DVB with core canonically isomorphic to E. It is easy to see that linear sections of $TE \to E$ are precisely linear vector fields (see the Appendix). Moreover, the inclusion $\Gamma(E) \to \Gamma_{\text{core}}(TE \to E)$ is the classical *vertical lift*, identifying a section of E with a fiber-wise constant vertical vector field on E itself. We will also call *core vector fields* the core sections of $TE \to E$.

The dual of TE over E is

$$\begin{array}{ccc} T^*E & \longrightarrow & E \\ \downarrow & & \downarrow \\ E^* & \longrightarrow & M \end{array}$$

We now pass to VB-algebroids.

Definition 1.2.5. A *VB-algebroid* is a DVB as in (1-12), equipped with a Lie algebroid structure $W \Rightarrow E$ such that the anchor $\rho_W : W \to TE$ is a vector bundle map covering a vector bundle map $\rho_A : A \to TM$ and the Lie bracket $[-,-]_W$ on sections of $W \to E$ satisfies

$$\begin{aligned} [\Gamma_{\text{lin}}(W,E), \Gamma_{\text{lin}}(W,E)]_W &\subset \Gamma_{\text{lin}}(W,E), \\ [\Gamma_{\text{lin}}(W,E), \Gamma_{\text{core}}(W,E)]_W &\subset \Gamma_{\text{core}}(W,E), \\ [\Gamma_{\text{core}}(W,E), \Gamma_{\text{core}}(W,E)]_W &= 0. \end{aligned} \tag{1-14}$$

Notice that, using the grading defined above, property (1-14) is equivalent to asking that the Lie bracket on $\Gamma(W,E)$ is of weight 0. This can be made very precise using the action of vector bundle automorphisms on multiderivations (see below).

Remark 1.2.6. Let $(W \Rightarrow E;\, A \Rightarrow M)$ be a VB-algebroid with core C, and let $(W^* \to C^*;\, A \to M)$ be its dual DVB. One can show that there is a canonical VB-algebroid structure $(W^*_A \Rightarrow C^*;\, A \Rightarrow M)$ on the latter, called the *dual VB-algebroid*. The dual VB-algebroid will appear only marginally in the sequel, so we do not discuss the details of this construction. For more information, see [Mackenzie 1998b].

Graded geometric description. There is a very useful description of VB-algebroids in terms of graded geometry. We begin discussing *linear vector fields* on (the total space of) a vector bundle $\mathcal{E} \to \mathcal{M}$ of graded manifolds. First we fix our notation. As already mentioned, a section ϕ of the dual bundle $\mathcal{E}^* \to \mathcal{M}$ determines a fiber-wise linear function ℓ_ϕ on $\mathcal{E}$. As in the nongraded case, a section ε of $\mathcal{E}$ itself determines a *fiber-wise constant* vector field $\varepsilon^\uparrow \in \mathfrak{X}(\mathcal{E})^\bullet$, its *vertical lift*, uniquely defined by

$$\varepsilon^\uparrow(\ell_\phi) := (-)^{|\varepsilon||\phi|}\langle \phi, \varepsilon\rangle.$$

We denote by $\mathfrak{X}_{\mathrm{core}}(\mathcal{E})^\bullet$ the space of core vector fields, i.e., fiber-wise constant vertical vector fields on $\mathcal{E}$. The correspondence $\varepsilon \mapsto \varepsilon^\uparrow$ establishes a graded $C^\infty(\mathcal{M})^\bullet$-module isomorphism $\Gamma(\mathcal{E})^\bullet \cong \mathfrak{X}_{\mathrm{core}}(\mathcal{E})^\bullet$. Now let $X \in \mathfrak{X}(\mathcal{E})^\bullet$. Then X is *linear* if it preserves fiber-wise linear functions. Equivalently, X is linear if the (graded) commutator $[X, -]$ preserves fiber-wise constant vector fields. We denote by $\mathfrak{X}_{\mathrm{lin}}(\mathcal{E})^\bullet$ the space of linear vector fields on $\mathcal{E}$. Notice that linear vector fields also preserve fiber-wise constant functions. Finally, similarly as in the nongraded case, denote by $\mathfrak{D}(\mathcal{E})^\bullet$ the space of graded derivations of $\mathcal{E}$. There is a canonical isomorphism of graded Lie algebras and graded $C^\infty(\mathcal{M})^\bullet$-modules $\mathfrak{X}_{\mathrm{lin}}(\mathcal{E})^\bullet \to \mathfrak{D}(\mathcal{E})^\bullet$, $X \mapsto D_X$, implicitly defined by $(D_X\varepsilon)^\uparrow = [X, \varepsilon^\uparrow]$, for all $\varepsilon \in \Gamma(\mathcal{E})^\bullet$.

Now, we have already recalled that Lie algebroids are equivalent to DG-manifolds concentrated in degree 0 and 1. For VB-algebroids we have an analogous result [Voronov 2012] that we now briefly explain. Recall that a DG-vector bundle is a vector bundle of graded manifolds $\mathcal{E} \to \mathcal{M}$ such that $\mathcal{E}$ and $\mathcal{M}$ are both DG-manifolds, with homological vector fields $Q_\mathcal{E}$ and $Q_\mathcal{M}$, respectively, and, additionally, $Q_\mathcal{E}$ is linear, and projects onto $Q_\mathcal{M}$. Equivalently $\mathcal{E} \to \mathcal{M}$ is a vector bundle of graded manifolds, $\mathcal{M}$ is a DG-manifold, with homological vector field $Q_\mathcal{M}$, $\mathcal{E}$ is equipped with a homological derivation $D_\mathcal{E}$, i.e., a degree 1 derivation such that $[D_\mathcal{E}, D_\mathcal{E}] = 0$, and, additionally, the symbol of $D_\mathcal{E}$ is precisely $Q_\mathcal{M}$. For more details about DG-vector bundles see, e.g., [Vitagliano 2016].

Finally, let $(W \to E;\, A \to M)$ be a DVB. If we shift the degree in the fibers of both $W \to E$ and $A \to M$ (and use the functoriality of the shift) we get a vector bundle of graded manifolds, denoted $W[1]_E \to A[1]$. If $(W \Rightarrow E;\, A \Rightarrow M)$ is a VB-algebroid, then $W[1]_E \to A[1]$ is a DG-vector bundle concentrated in degree 0 and 1.

Theorem 1.2.7 (see [Voronov 2012]). *The correspondence* $(W \Rightarrow E;\, A \Rightarrow M) \rightsquigarrow (W[1]_E \to A[1])$ *establishes an equivalence between the category of VB-algebroids and the category of DG-vector bundles concentrated in degree* 0 *and* 1.

1.3. ***The linear deformation complex of a VB-algebroid.*** In this subsection we introduce the main object of this paper: the *linear deformation complex of a VB-algebroid*, first introduced in [Esposito et al. 2016] (for different purposes). Actually, the whole discussion in Section 1.1 extends to VB-algebroids. We skip most of the proofs; they can be carried out in a very similar way as for plain Lie algebroids.

We begin with a DVB $(W \to E; A \to M)$. Denote by $\mathfrak{D}^\bullet(W, E)$ the space of multiderivations of the vector bundle $W \to E$. As in Section 1.2, denote by h the homogeneity structure of $W \to A$. The action of h induces a grading on the space of multiderivations.

Definition 1.3.1. A multiderivation $c \in \mathfrak{D}^\bullet(W, E)$ is *homogeneous of weight* q (or, simply, of *weight* q) if $h_\lambda^* c = \lambda^q c$ for every $\lambda > 0$. A multiderivation is *linear* if it is of weight 0, and it is *core* if it is of weight -1.

We denote by $\mathfrak{D}^\bullet_q(W, E)$ the space of multiderivations of weight q, and by $\mathfrak{D}^\bullet_{\mathrm{lin}}(W, E)$ and $\mathfrak{D}^\bullet_{\mathrm{core}}(W, E)$, respectively, the spaces of linear and core multiderivations.

As $\Gamma_{\mathrm{core}}(W, E)$ and $\Gamma_{\mathrm{lin}}(W, E)$ generate $\Gamma(W, E)$, a multiderivation is completely characterized by its action, and the action of its symbol, on linear and core sections. From (1-1) and the fact that there are no nonzero sections of weight less than -1, it then follows *there are no nonzero multiderivations of weight less than* -1. Moreover:

Proposition 1.3.2. *Let c be a k-derivation of $W \to E$. Then c is linear if and only if all the following conditions are satisfied*:

(1) $c(\tilde{\alpha}_1, \dots, \tilde{\alpha}_k)$ *is a linear section*,

(2) $c(\tilde{\alpha}_1, \dots, \tilde{\alpha}_{k-1}, \bar{\chi}_1)$ *is a core section*,

(3) $c(\tilde{\alpha}_1, \dots, \tilde{\alpha}_{k-i}, \bar{\chi}_1, \dots, \bar{\chi}_i) = 0$,

(4) $\sigma_c(\tilde{\alpha}_1, \dots, \tilde{\alpha}_{k-1})$ *is a linear vector field*,

(5) $\sigma_c(\tilde{\alpha}_1, \dots, \tilde{\alpha}_{k-2}, \bar{\chi}_1)$ *is a core vector field*,

(6) $\sigma_c(\tilde{\alpha}_1, \dots, \tilde{\alpha}_{k-i-1}, \bar{\chi}_1, \dots, \bar{\chi}_i) = 0$

for all linear sections $\tilde{\alpha}_1, \dots, \tilde{\alpha}_k$, all core sections $\bar{\chi}_1, \dots, \bar{\chi}_i$ of $W \to E$, and all $i \geq 2$.

Proof. This follows directly from Lemma 1.2.2; see also [Esposito et al. 2016]. □

In particular, a linear k-derivation is uniquely determined by its action on k linear sections and on $k-1$ linear sections and a core section, and by the action of its symbol on $k-1$ linear sections and on $k-2$ linear sections and a core section; see also [Esposito et al. 2016, Theorem 3.34].

It immediately follows from (1-2) that $\mathfrak{D}^\bullet_{\mathrm{lin}}(W, E)[1]$ is a graded Lie subalgebra of $\mathfrak{D}^\bullet(W, E)[1]$. The following proposition is then straightforward.

Proposition 1.3.3. *VB-algebroid structures on the DVB* $(W \to E;\ A \to M)$ *are in one-to-one correspondence with Maurer–Cartan elements in* $\mathfrak{D}^\bullet_{\mathrm{lin}}(W, E)[1]$.

Fix a VB-algebroid structure $(W \Rightarrow E;\ A \Rightarrow M)$ on the DVB $(W \to E;\ A \to M)$, and denote by b_W the Lie bracket on sections of $W \to E$. We also denote by $C^\bullet_{\mathrm{def}}(W, E)$ the deformation complex of the top algebroid $W \Rightarrow E$. It is clear that b_W is a linear biderivation of $W \to E$, i.e., $b_W \in \mathfrak{D}^2_{\mathrm{lin}}(W, E)$. Hence $\mathfrak{D}^\bullet_{\mathrm{lin}}(W, E)[1]$ is a sub-DGLA of $C^\bullet_{\mathrm{def}}(W, E)$, denoted $C^\bullet_{\mathrm{def,lin}}(W)$, and called the *linear deformation complex* of $W \Rightarrow E$. Its cohomology is denoted $H^\bullet_{\mathrm{def,lin}}(W)$ and called the *linear deformation cohomology* of $(W \Rightarrow E;\ A \Rightarrow M)$.

Definition 1.3.4. A *linear deformation* of b_W (or simply a *deformation*, if this does not lead to confusion) is a VB-algebroid structure on the DVB $(W \to E;\ A \to M)$.

Exactly as for Lie algebroids, Proposition 1.3.3 is equivalent to saying that *deformations of* b_W *are in one-to-one correspondence with Maurer–Cartan elements of* $C^\bullet_{\mathrm{def,lin}}(W)$.

Let b_0, b_1 be linear deformations of b_W. We say that b_0 and b_1 are *equivalent* if there exists a DVB isotopy taking b_0 to b_1, i.e., a smooth path of DVB automorphisms $\phi_t : W \to W$, $t \in [0, 1]$ such that $\phi_0 = \mathrm{id}_W$ and $\phi_1^* b_1 = b_0$. On the other hand, two Maurer–Cartan elements c_0, c_1 in $C^\bullet_{\mathrm{def,lin}}(W)$ are *gauge-equivalent* if they are interpolated by a smooth path of 1-cochains $c_t \in C^\bullet_{\mathrm{def,lin}}(W)$, and c_t is a solution of the ODE

$$\frac{dc_t}{dt} = \delta \Delta_t + [\![c_t, \Delta_t]\!]$$

for some smooth path of 0-cochains $\Delta_t \in C^\bullet_{\mathrm{def,lin}}(W)$, $t \in [0, 1]$. Equivalently,

$$\frac{db_t}{dt} = [\![b_t, \Delta_t]\!],$$

where $b_t = b_W + c_t$.

Proposition 1.3.5. *Deformations of the VB-algebroid* $(W \Rightarrow E;\ A \Rightarrow M)$ *are controlled by the DGLA* $C^\bullet_{\mathrm{def,lin}}(W)$ *in the following sense. Let* $b_0 = b_W + c_0$, $b_1 = b_W + c_1$ *be linear deformations of* b_W. *If* b_0, b_1 *are equivalent, then* c_0, c_1 *are gauge-equivalent. If* M *is compact, the converse is also true.*

Proof. The proof is similar to that of Proposition 1.1.5, with linear derivations replacing derivations and DVB automorphisms replacing vector bundle automorphisms. We only need to be careful when using the compactness hypothesis. Recall from [Esposito et al. 2016] that a linear derivation generates a flow by DVB automorphisms. In particular, if Δ_t is a time-dependent linear derivation of $W \to E$, then its symbol $X_t = \sigma(\Delta_t) \in \mathfrak{X}(E)$ is a *linear vector field*, hence it generates a flow by vector bundle automorphisms of E. From the compactness of M, it follows that X_t, hence the flow of Δ_t, is complete. □

Remark 1.3.6. An *infinitesimal deformation* of $(W \Rightarrow E; A \Rightarrow M)$ is an element $c \in C^1_{\mathrm{def,lin}}(W)$ such that $\delta c = 0$, i.e., c is a 1-cocycle in $C^\bullet_{\mathrm{def,lin}}(W)$. If c_t is a smooth path of Maurer–Cartan elements starting at 0, then $(dc_t/dt)|_{t=0}$ is an infinitesimal deformation of $(W \Rightarrow E; A \Rightarrow M)$. Similarly as for Lie algebroids, $H^1_{\mathrm{def,lin}}(W)$ is the formal tangent space to the moduli space of linear deformations under gauge equivalence. It also follows from standard deformation theory arguments that $H^2_{\mathrm{def,lin}}(W)$ contains obstructions to the extension of an infinitesimal linear deformation to a formal one. Finally, we interpret 0-degree deformation cohomologies. It easily follows from the definition that 0-cocycles in $C^\bullet_{\mathrm{def,lin}}(A)$ are *infinitesimal multiplicative* (IM) derivations of $(W \Rightarrow E; A \Rightarrow M)$ i.e., derivations of $W \to E$ generating a flow by VB-algebroid automorphisms [Esposito et al. 2016]. Among those, 1-cocycles are *inner IM derivations*, i.e., IM derivations of the form $[\tilde{\alpha}, -]$ for some *linear* section $\tilde{\alpha}$ of $W \to E$. So $H^0_{\mathrm{def,lin}}(W)$ consists of *outer IM derivations*. See [Esposito et al. 2016] for more details.

Alternative descriptions. Let $(W \Rightarrow E; A \Rightarrow M)$ be a VB-algebroid. Then $W \Rightarrow E$ is a Lie algebroid. As in the graded geometric description section we denote by $W[1]_E$ the DG-manifold obtained from W shifting by the degree in the fibers of $W \to E$. So $C^\bullet_{\mathrm{def}}(W) \cong \mathfrak{X}(W[1]_E)^\bullet$. Moreover, it is easy to see from (1-9) and Proposition 1.3.2 that a deformation cochain $c \in C^\bullet_{\mathrm{def}}(W)$ is linear if and only if the corresponding vector field $\delta_c \in \mathfrak{X}(W[1]_E)^\bullet$ is a linear vector field with respect to the vector bundle structure $W[1]_E \to A[1]$. So there is a canonical isomorphism of DGLAs

$$C^\bullet_{\mathrm{def,lin}}(W) \cong \mathfrak{X}_{\mathrm{lin}}(W[1]_E)^\bullet.$$

As linear vector fields are equivalent to derivations, we also get

$$C^\bullet_{\mathrm{def,lin}}(W) \cong \mathfrak{D}(W[1]_E, A[1])^\bullet \tag{1-15}$$

as DGLAs.

Deformations of A from linear deformations of W. There is a natural surjection $C^\bullet_{\mathrm{def,lin}}(W) \to C^\bullet_{\mathrm{def}}(A)$ which is easily described in the graded geometric picture: it is just the projection

$$\mathfrak{X}_{\mathrm{lin}}(W[1]_E)^\bullet \to \mathfrak{X}(A[1])^\bullet$$

of linear vector fields on the base. Equivalently, it is the symbol map

$$\sigma : \mathfrak{D}(W[1]_E, A[1])^\bullet \to \mathfrak{X}(A[1])^\bullet.$$

In particular, we get a short exact sequence of DGLAs

$$0 \to \mathfrak{End}(W[1]_E)^\bullet \to \mathfrak{X}_{\mathrm{lin}}(W[1]_E)^\bullet \to \mathfrak{X}(A[1])^\bullet \to 0, \tag{1-16}$$

where $\mathfrak{End}(W[1]_E)^\bullet$ is the space of (graded) endomorphisms of $W[1]_E \to A[1]$.

Equivalently, there is a short exact sequence

$$0 \to \mathfrak{End}(W[1]_E)^\bullet \to C^\bullet_{\mathrm{def,lin}}(W) \to C^\bullet_{\mathrm{def}}(A) \to 0. \tag{1-17}$$

Note that the sub-DGLA $\mathfrak{End}(W[1]_E)^\bullet$ *controls deformations of* $(W \Rightarrow E; A \Rightarrow M)$ *that fix* $A \Rightarrow M$, i.e., deformations of W that fix b_A (the Lie algebroid structure on A) identify with Maurer–Cartan elements in $\mathfrak{End}(W[1]_E)^\bullet$. Finally, we obtain a long exact sequence

$$\cdots \to H^k(\mathfrak{End}(W[1]_E)) \to H^k_{\mathrm{def,lin}}(W) \to H^k_{\mathrm{def}}(A) \to H^{k+1}(\mathfrak{End}(W[1]_E)) \to \cdots \tag{1-18}$$

connecting the linear deformation cohomology of W with the deformation cohomology of A.

Remark 1.3.7. A description of the subcomplex $\mathfrak{End}(W[1]_E)^\bullet \subset \mathfrak{X}_{\mathrm{lin}}(W[1]_E)^\bullet$ is not needed in terms of more classical data in this paper. However, we stress that this description exists in analogy with [Esposito et al. 2016, Theorem 3.34].

Deformations of the dual VB-algebroid. We conclude this section by noting that the linear deformation complex of a VB-algebroid is canonically isomorphic to that of its dual. Let $(W \Rightarrow E; A \Rightarrow M)$ be a VB-algebroid with core C, and let $(W^*_A \Rightarrow C^*; A \Rightarrow M)$ be the dual VB-algebroid.

Theorem 1.3.8. *There is a canonical isomorphism of DGLAs*

$$C^\bullet_{\mathrm{def,lin}}(W) \cong C^\bullet_{\mathrm{def,lin}}(W^*_A).$$

Proof. There is an easy proof exploiting graded geometry. We only sketch it, and leave the straightforward details to the reader. So, first of all, it is easy to see, e.g., in local coordinates, that the vector bundles of graded manifolds $W^*_A[1]_{C^*} \to A[1]$ and $W[1]^*_E \to A[1]$ are actually isomorphic up to a shift in the degree of the fiber coordinates. Additionally, derivations of a vector bundle of graded manifolds are canonically isomorphic to that of

(1) its dual,

(2) any vector bundle obtained from it by a shift in the degree of the fibers.

We conclude that

$$\begin{aligned} C^\bullet_{\mathrm{def,lin}}(W) &\cong \mathfrak{D}(W[1]_E, A[1])^\bullet \cong \mathfrak{D}(W[1]^*_E, A[1])^\bullet \\ &\cong \mathfrak{D}(W^*[1]_{C^*}, A[1])^\bullet \cong C^\bullet_{\mathrm{def,lin}}(W^*_A). \end{aligned}$$

□

1.4. ***From deformation cohomology to linear deformation cohomology.*** Let $(W \Rightarrow E; A \Rightarrow M)$ be a VB-algebroid. We have shown that deformations of the VB-algebroid structure are controlled by a sub-DGLA $C^\bullet_{\mathrm{def,lin}}(W)$ of the deformation complex $C^\bullet_{\mathrm{def}}(W)$ of the top Lie algebroid $W \Rightarrow E$. In the next section, we show that the inclusion $C^\bullet_{\mathrm{def,lin}}(W) \hookrightarrow C^\bullet_{\mathrm{def}}(W)$ induces an inclusion $H^\bullet_{\mathrm{def,lin}}(W) \hookrightarrow H^\bullet_{\mathrm{def}}(W)$

in cohomology. In particular, given an infinitesimal linear deformation that is trivial as infinitesimal deformation of the Lie algebroid $W \Rightarrow A$, i.e., it is connected to the zero deformation by an infinitesimal isotopy of vector bundle maps, then it is also trivial as infinitesimal linear deformation, i.e., it is also connected to the zero deformation by an infinitesimal isotopy of DVB maps.

The key idea is adapting to the present setting the "homogenization trick" of [Cabrera and Drummond 2017]. Let $E \to M$ be a vector bundle. In their paper, Cabrera and Drummond consider the following natural projections from $C^\infty(E)$ to its $C^\infty(M)$-submodules $C^\infty_q(E)$ (of *weight q homogeneous functions*):

$$\mathrm{pr}_q : C^\infty(E) \to C^\infty_q(E), \quad f \mapsto \frac{1}{q!}\frac{d^q}{d\lambda^q}|_{\lambda=0} h_\lambda^* f. \tag{1-19}$$

Notice that $\mathrm{pr}_q(f)$ is just the degree q part of the (fiber-wise) Taylor polynomial of f. In the following, we adopt the notations from the Appendix and denote

$$\begin{aligned} \mathrm{core} := \mathrm{pr}_0 : C^\infty(E) \to C^\infty_{\mathrm{core}}(E), \qquad f \mapsto f_{\mathrm{core}} = h_0^* f, \\ \mathrm{lin} := \mathrm{pr}_1 : C^\infty(E) \to C^\infty_{\mathrm{lin}}(E), \qquad f \mapsto f_{\mathrm{lin}} = \frac{d}{d\lambda}|_{\lambda=0} h_\lambda^* f, \end{aligned} \tag{1-20}$$

where $C^\infty_{\mathrm{core}}(E) := C^\infty_0(E)$, and $C^\infty_{\mathrm{lin}}(E) := C^\infty_1(M)$.

Formula (1-19) does not apply directly to multiderivations. To see why, let $(W \to E, A \to M)$ be a DVB, let h be the homogeneity structure of $W \to A$, and let $c \in \mathfrak{D}^\bullet(W, E)$. Then the curve $\lambda \mapsto h_\lambda^* c$ is not defined in 0. Actually, $\lambda = 0$ is a "pole of order 1" for $h_\lambda^* c$. More precisely, we have the following:

Proposition 1.4.1. *The limit*

$$\lim_{\lambda \to 0} \lambda \cdot h_\lambda^* c$$

exists and defines a core multiderivation c_{core}.

Proof. The existence of the limit can be shown in coordinates. Also, for every $\mu \neq 0$,

$$h_\mu^* c_{\mathrm{core}} = h_\mu^*(\lim_{\lambda \to 0} \lambda \cdot h_\lambda^* c) = \lim_{\lambda \to 0} \lambda \cdot h_\mu^* h_\lambda^* c = \lim_{\lambda \to 0} \mu^{-1}(\lambda\mu \cdot h_{\lambda\mu}^* c) = \mu^{-1} c_{\mathrm{core}}. \ \square$$

The next proposition can be proved in the same way.

Proposition 1.4.2. *The limit*

$$\lim_{\lambda \to 0} (h_\lambda^* c - \lambda^{-1} \cdot c_{\mathrm{core}})$$

exists and defines a linear multiderivation c_{lin}.

So far we have defined maps

$$\begin{aligned} \mathrm{core} : \mathfrak{D}^\bullet(W, E) \to \mathfrak{D}^\bullet_{\mathrm{core}}(W, E), \quad c \mapsto \lim_{\lambda \to 0} \lambda \cdot h_\lambda^* c \\ \mathrm{lin} : \mathfrak{D}^\bullet(W, E) \to \mathfrak{D}^\bullet_{\mathrm{lin}}(W, E), \quad c \mapsto \lim_{\lambda \to 0} (h_\lambda^* c - \lambda^{-1} \cdot c_{\mathrm{core}}) \end{aligned} \tag{1-21}$$

that split the inclusions in $\mathfrak{D}^\bullet(W, E)$. We call the latter the *linearization map*.

Remark 1.4.3. Once we have removed the singularity at 0, we can proceed as in (1-19) and define the projections on homogeneous multiderivation of positive weights $q > 0$:

$$\mathrm{pr}_q : \mathfrak{D}^\bullet(W, E) \to \mathfrak{D}^\bullet_q(W, E), \quad c \mapsto \frac{1}{q!}\frac{d^q}{d\lambda^q}|_{\lambda=0}(h^*_\lambda c - \lambda^{-1} \cdot c_{\mathrm{core}}).$$

Now, let $(W \Rightarrow E, A \Rightarrow M)$ be a VB-algebroid. Then we have a linearization map

$$\mathrm{lin} : C^\bullet_{\mathrm{def}}(W) \to C^\bullet_{\mathrm{def,lin}}(W).$$

Theorem 1.4.4 (linearization of deformation cochains). *The linearization map is a cochain map splitting the inclusion $C^\bullet_{\mathrm{def,lin}}(W) \hookrightarrow C^\bullet_{\mathrm{def}}(W)$. In particular there is a direct sum decomposition*

$$C^\bullet_{\mathrm{def}}(W) \cong C^\bullet_{\mathrm{def,lin}}(W) \oplus \ker(\mathrm{lin})^\bullet.$$

of cochain complexes. Hence, the inclusion of linear deformation cochains into deformation cochains induces an injection

$$H^\bullet_{\mathrm{def,lin}}(W) \hookrightarrow H^\bullet_{\mathrm{def}}(W). \tag{1-22}$$

Proof. We only have to prove that the linearization preserves the differential $\delta = [\![b_W, -]\!]$ (here, as usual b_W is the Lie bracket on sections of $W \Rightarrow E$). Using the fact that b_W is linear, we have that δ commutes with h^*_λ. From (1-4) it is obvious that δ preserves limits. So

$$(\delta c)_{\mathrm{core}} = \lim_{\lambda\to 0} \lambda \cdot h^*_\lambda(\delta c) = \lim_{\lambda\to 0} \lambda \cdot \delta(h^*_\lambda c) = \delta\Big(\lim_{\lambda\to 0} \lambda \cdot h^*_\lambda c\Big) = \delta c_{\mathrm{core}},$$

and

$$(\delta c)_{\mathrm{lin}} = \lim_{\lambda\to 0} (h^*_\lambda(\delta c) - \lambda^{-1}\delta(c_{\mathrm{core}})) = \delta\Big(\lim_{\lambda\to 0} (h^*_\lambda c - \lambda^{-1} \cdot c_{\mathrm{core}})\Big) = \delta c_{\mathrm{lin}},$$

as desired. □

The inclusion (1-22) can be used to transfer vanishing results from deformation cohomology of the Lie algebroid $W \Rightarrow E$ to the linear deformation cohomology of the VB-algebroid $(W \Rightarrow E; A \Rightarrow M)$. For example, if $H^0_{\mathrm{def}}(W) = 0$, every Lie algebroid derivation of $W \Rightarrow E$ is inner, and hence every IM derivation of the VB-algebroid W is inner. Similarly, if $W \Rightarrow E$ has no nontrivial infinitesimal deformations, so does $(W \Rightarrow E; A \Rightarrow M)$, and so on.

As a first example, consider a vector bundle $E \to M$. Then

$$\begin{array}{ccc} TE & \Longrightarrow & E \\ \downarrow & & \downarrow \\ TM & \Longrightarrow & M \end{array}$$

is a VB-algebroid.

Proposition 1.4.5. *The linear deformation cohomology of* $(TE \Rightarrow E; TM \Rightarrow M)$ *is trivial.*

Proof. From Theorem 1.4.4, $H^\bullet_{\mathrm{def,lin}}(TE)$ embeds into the deformation cohomology $H^\bullet_{\mathrm{def}}(TE)$ of the tangent algebroid $TE \Rightarrow E$ which is trivial; see, for example, [Crainic and Moerdijk 2008]. □

Other applications of Theorem 1.4.4 will be considered in Section 2.

Remember from Section 1.3 that a linear deformation cochain $c \in C^k_{\mathrm{def,lin}}(W)$ is completely determined by its action on k linear sections and on $k-1$ linear sections and a core section, and the action of its symbol on $k-1$ linear sections and on $k-2$ linear sections and a core section. We conclude this subsection providing a slightly more explicit description of the linearization map (1-21) in terms of these restricted actions.

Proposition 1.4.6. *Let* $c \in \mathfrak{D}^k(W; E)$. *Then* c_{lin} *is completely determined by the following identities*:

(1) $c_{\mathrm{lin}}(\tilde\alpha_1, \dots, \tilde\alpha_k) = c(\tilde\alpha_1, \dots, \tilde\alpha_k)_{\mathrm{lin}}$,

(2) $c_{\mathrm{lin}}(\tilde\alpha_1, \dots, \tilde\alpha_{k-1}, \bar\chi) = c(\tilde\alpha_1, \dots, \tilde\alpha_{k-1}, \bar\chi)_{\mathrm{core}}$,

(3) $\sigma_{c_{\mathrm{lin}}}(\tilde\alpha_1, \dots, \tilde\alpha_{k-1}) = \sigma_c(\tilde\alpha_1, \dots, \tilde\alpha_{k-1})_{\mathrm{lin}}$,

(4) $\sigma_{c_{\mathrm{lin}}}(\tilde\alpha_1, \dots, \tilde\alpha_{k-2}, \bar\chi) = \sigma_c(\tilde\alpha_1, \dots, \tilde\alpha_{k-2}, \bar\chi)_{\mathrm{core}}$

for all $\tilde\alpha_1, \dots, \tilde\alpha_k \in \Gamma_{\mathrm{lin}}(W, E)$, $\bar\chi \in \Gamma_{\mathrm{core}}(W, E)$.

Proof. We first compute

$$\begin{aligned} c_{\mathrm{core}}(\tilde\alpha_1, \dots, \tilde\alpha_k) &= \lim_{\lambda\to 0} \lambda \cdot (h_\lambda^* c)(\tilde\alpha_1, \dots, \tilde\alpha_k) \\ &= \lim_{\lambda\to 0} \lambda\, h_\lambda^*(c(\tilde\alpha_1, \dots, \tilde\alpha_k)) = c(\tilde\alpha_1, \dots, \tilde\alpha_k)_{\mathrm{core}}. \end{aligned}$$

Then

$$\begin{aligned} c_{\mathrm{lin}}(\tilde\alpha_1, \dots, \tilde\alpha_k) &= \lim_{\lambda\to 0} (h_\lambda^* c - \lambda^{-1} \cdot c_{\mathrm{core}})(\tilde\alpha_1, \dots, \tilde\alpha_k) \\ &= \lim_{\lambda\to 0} \big(h_\lambda^*(c(\tilde\alpha_1, \dots, \tilde\alpha_k)) - \lambda^{-1} c(\tilde\alpha_1, \dots, \tilde\alpha_k)_{\mathrm{core}}\big) \\ &= c(\tilde\alpha_1, \dots, \tilde\alpha_k)_{\mathrm{lin}}. \end{aligned}$$

Identity (2) can be proved in a similar way. To prove (3) first notice that

$$\sigma_{c_{\mathrm{core}}} = \sigma_{\lim_{\lambda\to 0} \lambda \cdot h_\lambda^* c} = \lim_{\lambda\to 0} \lambda \cdot \sigma_{h_\lambda^* c} = \lim_{\lambda\to 0} \lambda \cdot h_\lambda^* \sigma_c,$$

where we used (1-3). Hence

$$\begin{aligned} \sigma_{c_{\mathrm{core}}}(\tilde\alpha_1, \dots, \tilde\alpha_{k-1}) &= \lim_{\lambda\to 0} (\lambda \cdot h_\lambda^* \sigma_c)(\tilde\alpha_1, \dots, \tilde\alpha_{k-1}) \\ &= \lim_{\lambda\to 0} (\lambda \cdot h_\lambda^*(\sigma_c(\tilde\alpha_1, \dots, \tilde\alpha_{k-1}))) = \sigma_c(\tilde\alpha_1, \dots, \tilde\alpha_{k-1})_{\mathrm{core}}. \end{aligned}$$

Similarly,

$$\begin{aligned}\sigma_{c_{\mathrm{lin}}} &= \sigma_{\lim_{\lambda\to 0}(h_\lambda^* c - \lambda^{-1}\cdot c_{\mathrm{core}})}\\ &= \lim_{\lambda\to 0} \sigma_{h_\lambda^* c - \lambda^{-1}\cdot c_{\mathrm{core}}}\\ &= \lim_{\lambda\to 0} (h_\lambda^* \sigma_c - \lambda^{-1}\sigma_{c_{\mathrm{core}}}),\end{aligned}$$

hence

$$\begin{aligned}\sigma_{c_{\mathrm{lin}}}(\tilde\alpha_1,\dots,\tilde\alpha_{k-1}) &= \lim_{\lambda\to 0}\big((h_\lambda^*\sigma_c)(\tilde\alpha_1,\dots,\tilde\alpha_{k-1}) - \lambda^{-1}\sigma_{c_{\mathrm{core}}}(\tilde\alpha_1,\dots,\tilde\alpha_{k-1})\big)\\ &= \lim_{\lambda\to 0}\big(h_\lambda^*(\sigma_c(\tilde\alpha_1,\dots,\tilde\alpha_{k-1})) - \lambda^{-1}\sigma_c(\tilde\alpha_1,\dots,\tilde\alpha_{k-1})_{\mathrm{core}}\big)\\ &= \sigma_c(\tilde\alpha_1,\dots,\tilde\alpha_{k-1})_{\mathrm{lin}}.\end{aligned}$$

Identity (4) can be proved in a similar way. □

2. Examples and applications

In this section we provide several examples. Examples in Sections 2.1, 2.4 and 2.5 parallel the analogous examples in [Crainic and Moerdijk 2008], connecting our linear deformation cohomology to known cohomologies. Examples in Sections 2.2, 2.3 and 2.6 are specific to VB-algebroids.

2.1. ***VB-algebras.*** A *VB-algebra* is a *vector bundle object in the category of Lie algebras*. In other words, it is a VB-algebroid of the form

$$\begin{array}{ccc}\mathfrak{h} & \Longrightarrow & \{0\}\\ \downarrow & & \downarrow\\ \mathfrak{g} & \Longrightarrow & \{*\}\end{array}$$

In particular, $\mathfrak{h}$ and $\mathfrak{g}$ are Lie algebras. Now, let $C := \ker(\mathfrak{h}\to\mathfrak{g})$ be the core of $(\mathfrak{h}\Rightarrow\{0\};\mathfrak{g}\Rightarrow\{*\})$. It easily follows from the definition of VB-algebroid that

- C is a representation of $\mathfrak{g}$, and
- $\mathfrak{h} = \mathfrak{g}\ltimes C$ is the semidirect product Lie algebra,
- $\mathfrak{h} = \mathfrak{g}\ltimes C\to\mathfrak{g}$ is the projection onto the first factor.

Let $\operatorname{End} C$ denote endomorphisms of the vector space C. In the present case, the short exact sequence (1-16) reads

$$0\to C^\bullet(\mathfrak{g},\operatorname{End} C)\to C^\bullet_{\mathrm{def,lin}}(\mathfrak{h})\to C^\bullet_{\mathrm{def}}(\mathfrak{g})\to 0, \tag{2-1}$$

where $C^\bullet(\mathfrak{g},\operatorname{End} C) = \wedge^\bullet\mathfrak{g}^*\otimes\operatorname{End} C$ is the Chevalley–Eilenberg complex of $\mathfrak{g}$ with coefficients in the induced representation $\operatorname{End} C$, and $C^\bullet_{\mathrm{def}}(\mathfrak{g}) = (\wedge^\bullet\mathfrak{g}^*\otimes\mathfrak{g})[1]$ is the Chevalley–Eilenberg complex with coefficients in the adjoint representation.

From the classical theory of Nijenhuis and Richardson [1966; 1967a; 1967b], the latter controls deformations of $\mathfrak{g}$, while the former controls deformations of the representation of $\mathfrak{g}$ on C.

The sequence (2-1) has a natural splitting in the category of graded Lie algebras. Namely, there is an obvious graded Lie algebra map

$$C^\bullet_{\mathrm{def}}(\mathfrak{g}) \to C^\bullet_{\mathrm{def,lin}}(\mathfrak{h}), \quad c \mapsto \tilde{c}$$

given by

$$\tilde{c}(v_1 + \chi_1, \ldots, v_{k+1} + \chi_{k+1}) := c(v_1, \ldots, v_{k+1})$$

for all $c \in C^k_{\mathrm{def}}(\mathfrak{g}) = \wedge^{k+1}\mathfrak{g}^* \otimes \mathfrak{g}$, and all $v_i + \chi_i \in \mathfrak{h} = \mathfrak{g} \oplus C$, $i = 1, \ldots, k+1$. It is clear that the inclusion $C^\bullet_{\mathrm{def}}(\mathfrak{g}) \to C^\bullet_{\mathrm{def,lin}}(\mathfrak{h})$ splits the projection $C^\bullet_{\mathrm{def,lin}}(\mathfrak{h}) \to C^\bullet_{\mathrm{def}}(\mathfrak{g})$. Hence

$$C^\bullet_{\mathrm{def,lin}}(\mathfrak{h}) \cong C^\bullet(\mathfrak{g}, \operatorname{End} C) \oplus C^\bullet_{\mathrm{def}}(\mathfrak{g}). \tag{2-2}$$

as graded Lie algebras. However (2-2) is not a DGLA isomorphism. We now describe the differential δ in $C^\bullet_{\mathrm{def,lin}}(\mathfrak{h})$ in terms of the splitting (2-2). First of all, denote by $\theta : \mathfrak{g} \to \mathfrak{End}\, C$ the action of $\mathfrak{g}$ on C, and let

$$\Theta : \wedge^\bullet \mathfrak{g}^* \otimes \mathfrak{g} \to \wedge^\bullet \mathfrak{g}^* \otimes \operatorname{End} C,$$

be the map obtained from θ by extension of scalars. From the properties of the action, Θ is actually a cochain map

$$\Theta : C^\bullet_{\mathrm{def}}(\mathfrak{g})[-1] \to C^\bullet(\mathfrak{g}, \operatorname{End} C).$$

Finally, a direct computation reveals that the isomorphism (2-2) identifies the differential in $C^\bullet_{\mathrm{def,lin}}(\mathfrak{h})$ with that of the mapping cone (denoted by $\mathrm{Cone}(\Theta)$):

$$(C^\bullet_{\mathrm{def,lin}}(\mathfrak{h}), \delta) \cong \mathrm{Cone}(\Theta)$$

as cochain complexes. Notice that the long exact cohomology sequence of the mapping cone is just (1-18).

2.2. *LA-vector spaces.* An *LA vector space* is a *Lie algebroid object in the category of vector spaces*. In other words, it is a VB-algebroid of the form

$$\begin{array}{ccc} W & \Longrightarrow & E \\ \downarrow & & \downarrow \\ \{0\} & \Longrightarrow & \{*\} \end{array}$$

In particular, W and E are vector spaces. Now, let $C := \ker(W \to E)$ be the core of $(W \Rightarrow E; \{0\} \Rightarrow \{*\})$. It easily follows from the definition of VB-algebroid that W identifies canonically with the direct sum $C \oplus E$ and all the structure maps are

completely determined by a linear map $\partial : C \to E$. Specifically, sections of $W \to E$ are the same as smooth maps $E \to C$, and given a basis (C_I) of C, the Lie bracket on maps $E \to C$ is given by

$$[f, g] = f^I(\partial C_I)^\uparrow g - g^I(\partial C_I)^\uparrow f, \tag{2-3}$$

where $f = f^I C_I$ and $g = g^I C_I$. It follows that the anchor $\rho : W \to TE$ is given on sections by

$$\rho(f) = f^I(\partial C_I)^\uparrow. \tag{2-4}$$

Linear deformations of $(W \Rightarrow E; \{0\} \Rightarrow \{*\})$ are the same as deformations of ∂ as a linear map. Let us describe the linear deformation complex explicitly. As the bottom Lie algebroid is trivial, $C^\bullet_{\mathrm{def,lin}}(W)$ consists of graded endomorphisms $\mathrm{End}(C[1] \oplus E)^\bullet$ of the graded vector space $W[1]_E = C[1] \oplus E$. From (2-3) and (2-4) the differential δ in $\mathrm{End}(C[1] \oplus E)^\bullet$ is just the commutator with ∂, meaning that the deformation cohomology consists of homotopy classes of graded cochain maps $(C[1] \oplus E, \partial) \to (C[1] \oplus E, \partial)$. More explicitly, $(\mathrm{End}(C[1] \oplus E)^\bullet, \delta)$ is concentrated in degrees $-1, 0, 1$. Namely, it is

$$0 \to \mathrm{Hom}(E, C)[1] \xrightarrow{\delta_0} \mathrm{End}(C) \oplus \mathrm{End}(E) \xrightarrow{\delta_1} \mathrm{Hom}(C, E)[-1] \to 0,$$

where δ_0 and δ_1 are given by:

$$\delta_0\phi = (\phi \circ \partial, \partial \circ \phi),$$
$$\delta_1(\psi_C, \psi_E) = \partial \circ \psi_C - \psi_E \circ \partial,$$

where $\phi \in \mathrm{Hom}(E, C)$, $\psi_C \in \mathrm{End}(C)$ and $\psi_E \in \mathrm{End}(E)$. We immediately conclude

$$H^\bullet_{\mathrm{def,lin}}(W) = \mathrm{End}(\mathrm{coker}\,\partial \oplus \ker\partial[1])^\bullet,$$

that is,

$$H^{-1}_{\mathrm{def,lin}}(W) = \mathrm{Hom}(\mathrm{coker}\,\partial, \ker\partial),$$
$$H^0_{\mathrm{def,lin}}(W) = \mathrm{End}(\mathrm{coker}\,\partial) \oplus \mathrm{End}(\ker\partial),$$
$$H^1_{\mathrm{def,lin}}(W) = \mathrm{Hom}(\ker\partial, \mathrm{coker}\,\partial).$$

This shows, for instance, that infinitesimal deformations of a linear map $\partial : C \to E$ are all trivial if and only if ∂ is injective or surjective, as expected.

2.3. ***Tangent and cotangent VB-algebroids.*** Let $A \Rightarrow M$ be a Lie algebroid. Then $(TA \Rightarrow TM; A \Rightarrow M)$ is a VB-algebroid, called the *tangent VB-algebroid* of A. The structure maps of the Lie algebroid $TA \Rightarrow TM$ are defined as follows. First of all recall that $(TA \to TM; A \to M)$ is a DVB whose core is canonically isomorphic to A itself. In particular, any section α of A determines a core section $\bar{\alpha}$ of $TA \to TM$. A section α of A also determines a linear section $T\alpha$ of $TA \to TM$: its tangent map.

Denote by $\tau : TM \to M$ the projection. In the following, for a vector field $X \in \mathfrak{X}(M)$, we denote by $X_{\tan} \in \mathfrak{X}(TM)$ its *tangent lift*. By definition, the flow of $X_{\tan}$ is obtained from the flow of X by taking the tangent diffeomorphisms. Equivalently, $X_{\tan}$ is the (linear) vector field on TM uniquely determined by

$$X_{\tan}(\ell_{df}) = \ell_{dX(f)} \quad \text{and} \quad X_{\tan}(\tau^*) = \tau^* X(f) \tag{2-5}$$

for all $f \in C^\infty(M)$. Here ℓ_{df} is the fiber-wise linear function on TM corresponding to the 1-form df (viewed as a section of the dual bundle T^*M). Notice that (2-5) can be used to define the tangent lift of vector fields on a graded manifold. This will be useful below.

Now we come back to the tangent VB-algebroid $(TA \Rightarrow TM; A \Rightarrow M)$. The anchor $\rho_{TA} : TA \to TTM$ is determined by

$$\rho_{TA}(T\alpha) = \rho(\alpha)_{\tan}, \quad \rho_{TA}(\bar{\alpha}) = \rho(\alpha)^{\uparrow}, \tag{2-6}$$

and the bracket $[-,-]_{TA}$ in $\Gamma(TA, TM)$ is completely determined by:

$$[T\alpha, T\beta]_{TA} = T[\alpha, \beta], \quad [T\alpha, \bar{\beta}]_{TA} = \overline{[\alpha, \beta]}, \quad [\bar{\alpha}, \bar{\beta}]_{TA} = 0 \tag{2-7}$$

for all $\alpha, \beta \in \Gamma(A)$. The dual VB-algebroid $(T^*A \Rightarrow A^*; A \Rightarrow M)$ of the tangent VB-algebroid is called the *cotangent VB-algebroid*. We want to discuss the linear deformation cohomology of $(TA \Rightarrow TM; A \Rightarrow M)$ (hence of $(T^*A \Rightarrow A^*; A \Rightarrow M)$). We use the graded geometric description. Deformation cochains of $TA \Rightarrow TM$ are vector fields on the graded manifold $TA[1]_{TM}$ obtained from TA shifting by one the degree in the fibers of the vector bundle $TA \to TM$. Linear deformation cochains are vector fields that are linear with respect to the vector bundle structure $TA[1]_{TM} \to A[1]$.

Lemma 2.3.1. *Let $TA[1]$ be the tangent bundle of $A[1]$ and let $\tau : TA[1] \to A[1]$ be the projection. There is a canonical isomorphism of vector bundles of graded manifolds*

$$\begin{array}{ccc} TA[1]_{TM} & \xrightarrow{\ \iota\ } & TA[1] \\ & \searrow \quad \swarrow & \\ & A[1] & \end{array}$$

uniquely determined by the following condition:

$$\langle \iota^* \ell_{d\omega}, T\alpha_1 \wedge \cdots \wedge T\alpha_k \rangle = \ell_{d\langle \omega, \alpha_1 \wedge \cdots \wedge \alpha_k \rangle} \tag{2-8}$$

for all $\omega \in C^\infty(A[1])^\bullet = C^\bullet(A)$ of degree k, all sections $\alpha_1, \ldots, \alpha_k \in \Gamma(A)$, and all k. Additionally

$$\langle \iota^* \ell_{d\omega}, T\alpha_1 \wedge \cdots \wedge T\alpha_{k-1} \wedge \bar{\alpha}_k \rangle = \tau^* \langle \omega, \alpha_1 \wedge \cdots \wedge \alpha_k \rangle. \tag{2-9}$$

Formulas (2-8) and (2-9) require some explanations. The expression $d\omega$ on the left-hand side should be interpreted as a 1-form on $A[1]$, the de Rham differential of the function ω, and $\ell_{d\omega}$ is the associated fiber-wise linear function on $TA[1]$. The pull-back of $\ell_{d\omega}$ along ι^* is a function on $TA[1]_{TM}$, i.e., a $C^\infty(TM)$-valued, skew-symmetric multilinear map on sections of $TA \to TM$. The $T\alpha$ are the tangent maps $T\alpha : TM \to TA$ of the $\alpha : M \to A$. In particular they are linear sections of $TA \to TM$. The right-hand side of (2-8) is the fiber-wise linear function on TM corresponding to the 1-form $d\langle\omega, \alpha_1 \wedge \cdots \wedge \alpha_k\rangle$ on M. Here we interpret ω as a skew-symmetric multilinear map on sections of A.

Proof of Lemma 2.3.1. Let (x^i) be coordinates on M, let (u_α) be a local basis of $\Gamma(A)$, and let (u^α) be the associated fiber-wise linear coordinates on A. These data determine coordinates $(x^i, \tilde{u}^\alpha)$ on $A[1]$ in the obvious way. In particular the x^i have degree 0 and the $\tilde{u}^\alpha$ have degree 1. We also consider standard coordinates $(x^i, u^\alpha, \dot{x}^i, \dot{u}^\alpha)$ induced by (x^i, u^α) on TA. Notice that $(u^\alpha, \dot{u}^\alpha)$ are fiber-wise linear coordinates with respect to the vector bundle structure $TA \to TM$. More precisely, they are the fiber-wise linear coordinates associated to the local basis $(Tu_\alpha, \bar{u}_\alpha)$ of $\Gamma(TA, TM)$. Next we denote by $(x^i, \dot{x}^i, \tilde{u}^\alpha, \tilde{\dot{u}}^\alpha)$ the induced coordinates on $TA[1]_{TM}$. They have degree 0, 0, 1, 1 respectively. We denote by $(x^i, \tilde{u}^\alpha, X^i, \widetilde{U}^\alpha)$ the standard coordinates on $TA[1]$ induced by $(x^i, \tilde{u}^\alpha)$. Define ι by putting

$$\iota^* X^i = \dot{x}^i \quad \text{and} \quad \iota^* \widetilde{U}^\alpha = \tilde{\dot{u}}^\alpha.$$

A direct computation exploiting the appropriate transition maps reveals that ι is globally well defined. Now we prove (2-8). We work in coordinates. Take a degree k function $\omega = f_{\alpha_1 \cdots \alpha_k}(x)\tilde{u}^{\alpha_1} \cdots \tilde{u}^{\alpha_k}$ on $A[1]$. A direct computation shows that

$$\iota^* \ell_{d\omega} = \frac{\partial f_{\alpha_1 \cdots \alpha_k}}{\partial x^i} \tilde{u}^{\alpha_1} \cdots \tilde{u}^{\alpha_k} \dot{x}^i + k f_{\alpha_1 \cdots \alpha_k} \tilde{u}^{\alpha_1} \cdots \tilde{u}^{\alpha_{k-1}} \tilde{\dot{u}}^{\alpha_k}. \tag{2-10}$$

Now, let $\alpha_1, \ldots, \alpha_k \in \Gamma(A)$, and $a = 1, \ldots, k$. If α_a is locally given by $\alpha_a = g_a^\alpha(x) u_\alpha$, then

$$T\alpha_a = \frac{\partial g_a^\alpha}{\partial x^i} \dot{x}^i \bar{u}_\alpha + g_a^\alpha T u_\alpha,$$

and, from (2-10),

$$\begin{aligned} \langle \iota^* \ell_{d\omega}, T\alpha_1 \wedge \cdots \wedge T\alpha_k \rangle &= k! \left(\frac{\partial f_{\alpha_1 \cdots \alpha_k}}{\partial x^i} g_1^{\alpha_1} \cdots g_k^{\alpha_k} + f_{\alpha_1 \cdots \alpha_k} g_1^{\alpha_1} \cdots g_{k-1}^{\alpha_{k-1}} \frac{\partial g_k^{\alpha_k}}{\partial x^i} \right) \dot{x}^i \\ &= k! \frac{\partial}{\partial x^i} (f_{\alpha_1 \cdots \alpha_k} g_1^{\alpha_1} \cdots g_k^{\alpha_k}) \dot{x}^i = \ell_{d\langle \omega, \alpha_1 \wedge \cdots \wedge \alpha_k \rangle}. \end{aligned}$$

Identity (2-9) is proved in a similar way. To see that there is no other vector bundle isomorphism $\iota : TA[1]_{TM} \to TA[1]$ with the same property (2-8) notice that $X^i = \ell_{dx^i}$ and $\widetilde{U}^\alpha = \ell_{d\tilde{u}^\alpha}$. Now use (2-8) to show that $\iota^* X^i = \dot{x}^i$ and $\iota^* \widetilde{U}^\alpha = \tilde{\dot{u}}^\alpha$. □

In the following we will identify $TA[1]_{TM}$ with $TA[1]$ via the isomorphism ι of Lemma 2.3.1. Now, recall that $C^\bullet_{\mathrm{def,lin}}(TA) = \mathfrak{X}_{\mathrm{lin}}(TA[1])^\bullet$ fits in the short exact sequence of DGLAs:

$$0 \to \mathfrak{End}(TA[1])^\bullet \to \mathfrak{X}_{\mathrm{lin}}(TA[1])^\bullet \to \mathfrak{X}(A[1])^\bullet \to 0. \tag{2-11}$$

The tangent lift

$$\mathrm{tan} : \mathfrak{X}(A[1])^\bullet \hookrightarrow \mathfrak{X}(TA[1])^\bullet, \quad X \mapsto X_{\mathrm{tan}} \tag{2-12}$$

splits the sequence (2-11) in the category of DGLAs. As $\mathfrak{X}(A[1])^\bullet = C^\bullet_{\mathrm{def}}(A)$, we immediately have the following:

Proposition 2.3.2. *For every Lie algebroid $A \Rightarrow M$ there is a direct sum decomposition*

$$H^\bullet_{\mathrm{def,lin}}(TA) = H^\bullet_{\mathrm{def,lin}}(T^*A) = H^\bullet(\mathfrak{End}(TA[1])) \oplus H^\bullet_{\mathrm{def}}(A).$$

In the last part of the subsection we describe the inclusion (2-12) in terms of deformation cochains. This generalizes (2-6) and (2-7) to possibly higher cochains. Using the canonical isomorphisms $C^\bullet_{\mathrm{def,lin}}(TA) = \mathfrak{X}_{\mathrm{lin}}(TA[1])^\bullet$, and $C^\bullet_{\mathrm{def}}(A) = \mathfrak{X}(A[1])^\bullet$ we get an inclusion

$$\mathrm{tan} : C^\bullet_{\mathrm{def}}(A) \hookrightarrow C^\bullet_{\mathrm{def,lin}}(TA), \quad c \mapsto c_{\mathrm{tan}}.$$

Proposition 2.3.3. *Let $c \in C^{k-1}_{\mathrm{def}}(A)$. Then $c_{\mathrm{tan}} \in C^{k-1}_{\mathrm{def,lin}}(TA)$ satisfies*:

(1) $c_{\mathrm{tan}}(T\alpha_1, \ldots, T\alpha_k) = Tc(\alpha_1, \ldots, \alpha_k)$,

(2) $c_{\mathrm{tan}}(T\alpha_1, \ldots, T\alpha_{k-1}, \bar{\alpha}_k) = \overline{c(\alpha_1, \ldots, \alpha_k)}$,

(3) $\sigma_{c_{\mathrm{tan}}}(T\alpha_1, \ldots, T\alpha_{k-1}) = \sigma_c(\alpha_1, \ldots, \alpha_{k-1})_{\mathrm{tan}}$,

(4) $\sigma_{c_{\mathrm{tan}}}(T\alpha_1, \ldots, T\alpha_{k-2}, \bar{\alpha}_{k-1}) = \sigma_c(\alpha_1, \ldots, \alpha_{k-1})^\uparrow$

for all $\alpha_1, \ldots, \alpha_k \in \Gamma(A)$. Identities (1)–(4) (*together with the fact that c_{tan} is a linear cochain*) *determine c_{tan} completely.*

Proof. We begin with (3). Recall that the tangent lift X_{tan} of a vector field X is completely determined by (2-5) (and this remains true in the graded setting). So, let $X \in \mathfrak{X}(A[1])^\bullet$ be the graded vector field corresponding to c (hence $X_{\mathrm{tan}} \in \mathfrak{X}(TA[1]_{TM})$ is the graded vector field corresponding to c_{tan}), let $f \in C^\infty(M)$ and let $\alpha_1, \ldots, \alpha_{k-1} \in \Gamma(A)$. Using (1-10), compute

$$\begin{aligned}\sigma_{c_{\mathrm{tan}}}(T\alpha_1, \ldots, T\alpha_{k-1})\ell_{df} &= \langle X_{\mathrm{tan}}(\ell_{df}), T\alpha_1 \wedge \cdots \wedge T\alpha_{k-1}\rangle \\ &= \langle \ell_{d(X(f))}, T\alpha_1 \wedge \cdots \wedge T\alpha_{k-1}\rangle.\end{aligned}$$

From (2-8),

$$\begin{aligned}\langle \ell_{d(X(f))}, T\alpha_1 \wedge \cdots \wedge T\alpha_{k-1}\rangle &= \ell_{d\langle X(f), \alpha_1 \wedge \cdots \wedge \alpha_{k-1}\rangle} = \ell_{d(\sigma_c(\alpha_1, \ldots, \alpha_{k-1})f)} \\ &= \sigma_c(\alpha_1, \ldots, \alpha_{k-1})_{\mathrm{tan}}\ell_{df}.\end{aligned}$$

Since $\sigma_{c_{\tan}}(T\alpha_1, \ldots, T\alpha_{k-1})$ and $\sigma_c(\alpha_1, \ldots, \alpha_{k-1})_{\tan}$ are both linear and project onto $\sigma_c(\alpha_1, \ldots, \alpha_{k-1})$, this is enough to conclude that $\sigma_{c_{\tan}}(T\alpha_1, \ldots, T\alpha_{k-1}) = \sigma_c(\alpha_1, \ldots, \alpha_{k-1})_{\tan}$. Identity (4) can be proved in a similar way using (2-9) and

$$\sigma_c(\alpha_1, \ldots, \alpha_{k-1})^{\uparrow}\ell_{df} = \tau^*\langle df, \sigma_c(\alpha_1, \ldots, \alpha_{k-1})\rangle = \tau^*(\sigma_c(\alpha_1, \ldots, \alpha_{k-1})f).$$

We now prove (1). Both sides of the identity are linear sections of $TA \to TM$ and one can easily check in local coordinates that a linear section $\tilde{\alpha}$ is completely determined by pairings of the form $\langle \ell_{d\varphi}, \tilde{\alpha}\rangle$. Here, φ is a section of $A^* \to M$ seen as a degree 1 function on $A[1]$, $d\varphi$ is its de Rham differential, and $\ell_{d\varphi}$ is the associated degree 1 fiber-wise linear function on $TA[1]$, which, in turn, can be interpreted as a 1-form on the algebroid $TA \Rightarrow TM$, as in Lemma 2.3.1.

So, take $\varphi \in \Gamma(A^*)$, $c \in C^{k-1}_{\mathrm{def}}(A)$, $\alpha_1, \ldots, \alpha_k \in \Gamma(A)$, and compute

$$\begin{aligned}
&\langle \ell_{d\varphi}, c_{\tan}(T\alpha_1, \ldots, T\alpha_k)\rangle \\
&= \langle X_{\mathrm{tot}}(\ell_{d\varphi}), T\alpha_1 \wedge \cdots \wedge T\alpha_k\rangle - \sum_i (-)^{k-i} \sigma_{c_{\tan}}(T\alpha_1, \ldots, \widehat{T\alpha_i}, \ldots, T\alpha_k)\langle \ell_{d\varphi}, T\alpha_i\rangle \\
&= \langle \ell_{d(X(\varphi))}, T\alpha_1 \wedge \cdots \wedge T\alpha_k\rangle - \sum_i (-)^{k-i} \sigma_c(\alpha_1, \ldots, \widehat{\alpha_i}, \ldots, \alpha_k)_{\tan}\ell_{d\langle \omega, \alpha_i\rangle} \\
&= \ell_{d\langle X(\varphi), \alpha_1 \wedge \cdots \wedge \alpha_k\rangle} - \sum_i (-)^{k-i} \ell_{d(\sigma_c(\alpha_1, \ldots, \widehat{\alpha_i}, \ldots, \alpha_k)\langle \omega, \alpha_i\rangle)} \\
&= \ell_{d\langle \varphi, c(\alpha_1, \ldots, \alpha_k)\rangle} = \langle \ell_{d\varphi}, Tc(\alpha_1, \ldots, \alpha_k)\rangle,
\end{aligned}$$

where we used, in particular, (1-9), the first equality in (2-5), identity (3), and (2-8). So (1) holds.

Identity (2) can be proved in a similar way using (1-9), both identities (2-5), identity (4), and (2-9). We leave to the reader the straightforward details. □

Remark 2.3.4. Proposition 2.3.3 shows, in particular, that the Lie bracket b_A on $\Gamma(A)$ and the Lie bracket b_{TA} in $\Gamma(TA, TM)$ are related by $b_{TA} = (b_A)_{\tan}$.

2.4. ***Partial connections.*** Let M be a manifold, $D \subset TM$ an involutive distribution, and let $\mathcal{F}$ be the integral foliation of D. In particular $D \Rightarrow M$ is a Lie algebroid with injective anchor. A flat (partial) D-connection ∇ in a vector bundle $E \to M$ defines a VB-algebroid

$$\begin{array}{ccc} H & \Longrightarrow & E \\ \downarrow & & \downarrow \\ D & \Longrightarrow & M \end{array}$$

where $H \subset TE$ is the *horizontal distribution* determined by D. Notice that the core of $(H \Rightarrow E;\ D \Rightarrow M)$ is trivial, and every VB-algebroid with injective (base) anchor and trivial core arises in this way. Hence, (small) deformations of $(H \Rightarrow E;\ D \Rightarrow M)$ are the same as simultaneous deformations of the foliation $\mathcal{F}$ and the flat partial

connection ∇. We now discuss the linear deformation cohomology. Denote by $q : E \to M$ the projection. First of all, the de Rham complex of $D \Rightarrow M$ is the same as leaf-wise differential forms $\Omega^\bullet(\mathcal{F})$ with the leaf-wise de Rham differential $d_{\mathcal{F}}$. Hence, the deformation complex of D consists of derivations of $\Omega^\bullet(\mathcal{F})$ (the differential being the graded commutator with $d_{\mathcal{F}}$). As the core of $(H \Rightarrow E; D \Rightarrow M)$ is trivial, there is a canonical isomorphism $H \cong q^*D$ of vector bundles over E. It easily follows that the linear deformation complex $(C^\bullet_{\mathrm{def,lin}}(H), \delta)$ consists of derivations of the graded module $\Omega^\bullet(\mathcal{F}, E)$ of E-valued, leaf-wise differential forms, and the differential δ is the commutator with the (leaf-wise partial) connection differential $d^\nabla_{\mathcal{F}}$. The kernel of $C^\bullet_{\mathrm{def,lin}}(H) \to C^\bullet_{\mathrm{def}}(D)$ consists of graded $\Omega^\bullet(\mathcal{F})$-linear endomorphisms of $\Omega^\bullet(\mathcal{F}, E)$. The latter are the same as $\operatorname{End} E$-valued leaf-wise differential forms $\Omega^\bullet(\mathcal{F}, \operatorname{End} E)$, and the restricted differential is the connection differential (corresponding to the induced connection in $\operatorname{End} E$).

Now, denote by $\nu = TM/D$ the normal bundle to $\mathcal{F}$. It is canonically equipped with the Bott connection ∇^{Bott}, and there is a deformation retraction, hence a quasi-isomorphism, $\pi : C^\bullet_{\mathrm{def}}(D) \to \Omega^\bullet(\mathcal{F}, \nu)$ that maps a deformation cochain c to the composition $\pi(c)$ of the symbol $\sigma_c : \wedge^\bullet D \to TM$ followed by the projection $TM \to \nu$. A similar construction can be applied to linear deformation cochains. To see this, first notice that derivations of E modulo covariant derivatives along ∇, $\mathfrak{D}(E)/\operatorname{im}\nabla$, are sections of a vector bundle $\tilde{\nu} \to M$. Additionally, $\tilde{\nu}$ is canonically equipped with a flat partial connection, also called the *Bott connection* and denoted ∇^{Bott}, defined by

$$\nabla^{\mathrm{Bott}}_X(\Delta \bmod \operatorname{im}\nabla) = [\nabla_X, \Delta] \bmod \operatorname{im}\nabla$$

for all $\Delta \in \mathfrak{D}(E)$, and $X \in \Gamma(D)$. The symbol map $\sigma : \mathfrak{D}(E) \to \mathfrak{X}(M)$ descends to a surjective vector bundle map $\tilde{\nu} \to \nu$, intertwining the Bott connections. As $\operatorname{End} E \cap \operatorname{im}\nabla = 0$, we have $\ker(\tilde{\nu} \to \nu) = \operatorname{End} E$. In other words, there is a short exact sequence of vector bundles with partial connections:

$$0 \to \operatorname{End} E \to \tilde{\nu} \to \nu \to 0.$$

Now, we define a surjective cochain map $\tilde{\pi} : C^\bullet_{\mathrm{def,lin}}(H) \to \Omega^\bullet(\mathcal{F}, \tilde{\nu})$. Let $\tilde{c}$ be a linear deformation cochain. Its symbol $\sigma_{\tilde{c}}$ maps linear sections of $H \to E$ to linear vector field on E. As $H \cong q^*D$, linear sections identify with plain sections of D. Accordingly $\sigma_{\tilde{c}}$ can be seen as a $\mathfrak{D}(E)$-valued D-form. Take this point of view and denote by $\tilde{\pi}(\tilde{c}) : \wedge^\bullet D \to \tilde{\nu}$ the composition of $\sigma_{\tilde{c}}$ followed by the projection $\mathfrak{D}(E) \to \Gamma(\tilde{\nu})$.

Summarizing, we have the commutative diagram

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \Omega^\bullet(\mathcal{F}, \operatorname{End} E) & \longrightarrow & C^\bullet_{\mathrm{def,lin}}(H) & \longrightarrow & C^\bullet_{\mathrm{def}}(D) & \longrightarrow & 0 \\
 & & \| & & \downarrow{\scriptstyle \tilde{\pi}} & & \downarrow{\scriptstyle \pi} & & \\
0 & \longrightarrow & \Omega^\bullet(\mathcal{F}, \operatorname{End} E) & \longrightarrow & \Omega^\bullet(\mathcal{F}, \tilde{\nu}) & \longrightarrow & \Omega^\bullet(\mathcal{F}, \nu) & \longrightarrow & 0
\end{array}$$

The rows are short exact sequences of DG-modules, and the vertical arrows are DG-module surjections. Additionally, π is a quasi-isomorphism. Hence, it immediately follows from the snake lemma and the five lemma that $\tilde{\pi}$ is a quasi-isomorphism as well. We have thus proved the following:

Proposition 2.4.1. *There is a canonical isomorphism of graded vector spaces between the linear deformation cohomology of the VB-algebroid* $(H \Rightarrow E; D \Rightarrow M)$, *and the leaf-wise cohomology with coefficients in* $\tilde{\nu}$:

$$H^\bullet_{\mathrm{def,lin}}(H) = H^\bullet(\mathcal{F}, \tilde{\nu}).$$

2.5. ***Lie algebra actions on vector bundles.*** Let $\mathfrak{g}$ be a (finite-dimensional, real) Lie algebra acting on a vector bundle $E \to M$ by infinitesimal vector bundle automorphisms. In particular $\mathfrak{g}$ acts on M and there is an associated action Lie algebroid $\mathfrak{g} \ltimes M \Rightarrow M$. Additionally, $\mathfrak{g}$ acts on the total space E by linear vector fields. Equivalently, there is a Lie algebra homomorphism $\mathfrak{g} \to \mathfrak{D}(E)$ covering the (infinitesimal) action $\mathfrak{g} \to \mathfrak{X}(M)$. It follows that $(\mathfrak{g} \ltimes E \Rightarrow E; \mathfrak{g} \ltimes M \Rightarrow M)$ is a VB-algebroid. We want to discuss linear deformation cohomologies of $\mathfrak{g} \ltimes E \Rightarrow E$. We begin reviewing remarks by Crainic and Moerdijk [2008] on the deformation cohomology of $\mathfrak{g} \ltimes M \Rightarrow M$ providing a graded geometric interpretation. The deformation complex $C^\bullet_{\mathrm{def}}(\mathfrak{g} \ltimes M)$ consists of vector fields on $(\mathfrak{g} \times M)[1] = \mathfrak{g}[1] \times M$. Denote by

$$\pi_{\mathfrak{g}} : \mathfrak{g}[1] \times M \to \mathfrak{g}[1]$$

the projection. Composition on the right with the pull-back

$$\pi_{\mathfrak{g}}^* : C^\infty(\mathfrak{g}[1])^\bullet \to C^\infty(\mathfrak{g}[1] \times M)^\bullet$$

establishes a projection from vector fields on $\mathfrak{g}[1] \times M$ to $\pi_{\mathfrak{g}}$-relative vector fields $\mathfrak{X}_{\mathrm{rel}}(\pi_{\mathfrak{g}})^\bullet$, i.e., vector fields on $\mathfrak{g}[1]$ with coefficients in functions on $\mathfrak{g}[1] \times M$:

$$\mathfrak{X}(\mathfrak{g}[1] \times M)^\bullet \to \mathfrak{X}_{\mathrm{rel}}(\pi_{\mathfrak{g}})^\bullet, \quad X \mapsto X \circ \pi_{\mathfrak{g}}^*. \tag{2-13}$$

The kernel of projection (2-13) consists of $\pi_{\mathfrak{g}}$-vertical vector fields $\mathfrak{X}^{\pi_{\mathfrak{g}}}(\mathfrak{g}[1] \times M)^\bullet$. Denote by $d_{\mathfrak{g}} \in \mathfrak{X}(\mathfrak{g}[1] \times M)^\bullet$ the homological vector field on $\mathfrak{g}[1] \times M$. The graded commutator $\delta := [d_{\mathfrak{g}}, -]$ preserves $\pi_{\mathfrak{g}}$-vertical vector fields. Hence there is a short exact sequence of cochain complexes:

$$0 \to \mathfrak{X}^{\pi_{\mathfrak{g}}}(\mathfrak{g}[1] \times M)^\bullet \to \mathfrak{X}(\mathfrak{g}[1] \times M)^\bullet \to \mathfrak{X}_{\mathrm{rel}}(\pi_{\mathfrak{g}})^\bullet \to 0. \tag{2-14}$$

Now, $\mathfrak{X}(\mathfrak{g}[1] \times M)^\bullet$ is exactly the deformation complex of $\mathfrak{g} \ltimes M$. Similarly, $\mathfrak{X}_{\mathrm{rel}}(\pi_{\mathfrak{g}})^\bullet$ is (canonically isomorphic to) the Chevalley–Eilenberg cochain complex of $\mathfrak{g}$ with coefficients in $C^\infty(M) \otimes \mathfrak{g}$, the tensor product of $C^\infty(M)$ and the adjoint representation, up to a shift by 1. Following [Crainic and Moerdijk 2008], we shortly denote this tensor product by $\mathfrak{g}_M$. Finally, $\mathfrak{X}^{\pi_{\mathfrak{g}}}(\mathfrak{g}[1] \times M)^\bullet$ is canonically

isomorphic to the Chevalley–Eilenberg cochain complex of $\mathfrak{g}$ with coefficients in $\mathfrak{X}(M)$. So there is a short exact sequence of cochain complexes

$$0 \to C^\bullet(\mathfrak{g}, \mathfrak{X}(M)) \to C^\bullet_{\mathrm{def}}(\mathfrak{g} \ltimes M) \to C^{\bullet+1}(\mathfrak{g}, \mathfrak{g}_M) \to 0,$$

and a long exact cohomology sequence

$$\cdots \to H^k(\mathfrak{g}, \mathfrak{X}(M)) \to H^k_{\mathrm{def}}(\mathfrak{g} \ltimes M) \to H^{k+1}(\mathfrak{g}, \mathfrak{g}_M) \to H^{k+1}(\mathfrak{g}, \mathfrak{X}(M)) \to \cdots \tag{2-15}$$

We pass to $\mathfrak{g} \ltimes E$. The linear deformation complex $C^\bullet_{\mathrm{def,lin}}(\mathfrak{g} \ltimes E)$ consists of linear vector fields on $\mathfrak{g}[1] \times E$. As above, we consider the projection $\tilde{\pi}_\mathfrak{g} : \mathfrak{g}[1] \times E \to \mathfrak{g}[1]$. Composition on the right with the pull-back $\tilde{\pi}_\mathfrak{g}^*$ establishes a projection:

$$\mathfrak{X}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet \to \mathfrak{X}_{\mathrm{rel}}(\pi_\mathfrak{g})^\bullet, \quad X \mapsto X \circ \tilde{\pi}_\mathfrak{g}^*$$

(beware, the range consists of $\pi_\mathfrak{g}$-relative, not $\tilde{\pi}_\mathfrak{g}$-relative, vector fields) whose kernel consists of $\tilde{\pi}_\mathfrak{g}$-vertical linear vector fields $\mathfrak{X}^{\pi_\mathfrak{g}}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet$. Hence there is a short exact sequence of cochain complexes:

$$0 \to \mathfrak{X}^{\tilde{\pi}_\mathfrak{g}}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet \to \mathfrak{X}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet \to \mathfrak{X}_{\mathrm{rel}}(\pi_\mathfrak{g})^\bullet \to 0. \tag{2-16}$$

Using the projection $\mathfrak{X}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet \to \mathfrak{X}(\mathfrak{g}[1] \times M)^\bullet$, we can combine sequences (2-16) and (2-14) in an exact diagram

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & & & \\
 & & \downarrow & & \downarrow & & & & \\
0 & \longrightarrow & \mathfrak{End}(\mathfrak{g}[1] \times E)^\bullet & = & \mathfrak{End}(\mathfrak{g}[1] \times E)^\bullet & \longrightarrow & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & \mathfrak{X}^{\tilde{\pi}_\mathfrak{g}}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet & \longrightarrow & \mathfrak{X}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet & \longrightarrow & \mathfrak{X}_{\mathrm{rel}}(\pi_\mathfrak{g})^\bullet & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \| & & \\
0 & \longrightarrow & \mathfrak{X}^{\pi_\mathfrak{g}}(\mathfrak{g}[1] \times M)^\bullet & \longrightarrow & \mathfrak{X}(\mathfrak{g}[1] \times M)^\bullet & \longrightarrow & \mathfrak{X}_{\mathrm{rel}}(\pi_\mathfrak{g})^\bullet & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

where, as usual, $\mathfrak{End}(\mathfrak{g}[1] \times E)^\bullet$ consists of graded endomorphisms of the vector bundle $\mathfrak{g}[1] \times E \to \mathfrak{g}[1] \times M$ (covering the identity). Now, $\mathfrak{X}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet$ is the linear deformation complex of $(\mathfrak{g} \ltimes E \Rightarrow E; \mathfrak{g} \ltimes M \Rightarrow M)$, and $\mathfrak{End}(\mathfrak{g}[1] \times E)^\bullet$ is canonically isomorphic to the Chevalley–Eilenberg cochain complex of $\mathfrak{g}$ with coefficients in $\mathfrak{End}\, E$, endomorphisms of E (covering the identity). Finally, $\mathfrak{X}^{\tilde{\pi}_\mathfrak{g}}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet$ is canonically isomorphic to the Chevalley–Eilenberg cochain complex of $\mathfrak{g}$ with coefficients in $\mathfrak{D}(E)$. The isomorphism

$$C^\bullet(\mathfrak{g}, \mathfrak{D}(E)) \xrightarrow{\cong} \mathfrak{X}^{\tilde{\pi}_\mathfrak{g}}_{\mathrm{lin}}(\mathfrak{g}[1] \times E)^\bullet$$

maps a cochain $\omega \otimes \Delta$ to the vector field $\tilde{\pi}_\mathfrak{g}^*(f_\omega) X_\Delta$, where f_ω is the function on $\mathfrak{g}[1]$ corresponding to $\omega \in C^\bullet(\mathfrak{g})$, and X_Δ is the unique $\tilde{\pi}_\mathfrak{g}$-vertical vector field on $\mathfrak{g}[1] \times E$ projecting on the linear vector field on E corresponding to derivation Δ.

We conclude that there is an exact diagram of cochain complexes

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & & & \\
 & & \downarrow & & \downarrow & & & & \\
0 & \longrightarrow & C^\bullet(\mathfrak{g}, \mathfrak{End}\, E) & = & C^\bullet(\mathfrak{g}, \mathfrak{End}\, E) & \longrightarrow & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & C^\bullet(\mathfrak{g}, \mathfrak{D}(E)) & \longrightarrow & C^\bullet_{\mathrm{def,lin}}(\mathfrak{g} \ltimes E) & \longrightarrow & C^{\bullet+1}(\mathfrak{g}, \mathfrak{g}_M) & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \| & & \\
0 & \longrightarrow & C^\bullet(\mathfrak{g}, \mathfrak{X}(M)) & \longrightarrow & C^\bullet_{\mathrm{def}}(\mathfrak{g} \ltimes M) & \longrightarrow & C^{\bullet+1}(\mathfrak{g}, \mathfrak{g}_M) & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

This proves the following:

Proposition 2.5.1. *Let $\mathfrak{g}$ be a Lie algebra acting on a vector bundle $E \to M$ by infinitesimal vector bundle automorphisms. The linear deformation cohomology of the VB-algebroid $(\mathfrak{g} \ltimes E \Rightarrow E, \mathfrak{g} \ltimes M \Rightarrow M)$ fits in the exact diagram*

$$\begin{array}{ccccccccc}
 & & \vdots & & \vdots & & \vdots & & \vdots \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
\cdots & \longrightarrow & H^k(\mathfrak{g}, \mathfrak{End}\, E) & = & H^k(\mathfrak{g}, \mathfrak{End}\, E) & \longrightarrow & 0 & \longrightarrow & H^{k+1}(\mathfrak{g}, \mathfrak{End}\, E) \longrightarrow \cdots \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
\cdots & \longrightarrow & H^k(\mathfrak{g}, \mathfrak{D}(E)) & \longrightarrow & H^k_{\mathrm{def,lin}}(\mathfrak{g} \ltimes E) & \longrightarrow & H^{k+1}(\mathfrak{g}, \mathfrak{g}_M) & \longrightarrow & H^{k+1}(\mathfrak{g}, \mathfrak{D}(E)) \longrightarrow \cdots \\
 & & \downarrow & & \downarrow & & \| & & \downarrow \\
\cdots & \longrightarrow & H^k(\mathfrak{g}, \mathfrak{X}(M)) & \longrightarrow & H^k_{\mathrm{def}}(\mathfrak{g} \ltimes M) & \longrightarrow & H^{k+1}(\mathfrak{g}, \mathfrak{g}_M) & \longrightarrow & H^{k+1}(\mathfrak{g}, \mathfrak{X}(M)) \longrightarrow \cdots \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
\cdots & \longrightarrow & H^{k+1}(\mathfrak{g}, \mathfrak{End}\, E) & = & H^{k+1}(\mathfrak{g}, \mathfrak{End}\, E) & \longrightarrow & 0 & \longrightarrow & H^{k+2}(\mathfrak{g}, \mathfrak{End}\, E) \longrightarrow \cdots \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
 & & \vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

2.6. ***Type 1 VB-algebroids.*** Let $(W \Rightarrow E; A \Rightarrow M)$ be a VB-algebroid with core C. The *core-anchor* of $(W \Rightarrow E; A \Rightarrow M)$ is the vector bundle map $\partial : C \to E$ defined as follows. Let χ be a section of C, and let $\overline{\chi}$ be the corresponding core section of $W \to E$. The anchor $\rho : W \to TE$ maps $\overline{\chi}$ to a core vector field $\rho(\overline{\chi})$ on E. In turn $\rho(\overline{\chi})$ is the vertical lift of a section ε of E. By definition, $\partial\chi = \varepsilon$.

According to a definition by Gracia-Saz and Mehta [2010], a VB-algebroid is *type* 1 (resp. *type* 0) if the core-anchor is an isomorphism (resp. is the zero map). More generally, $(W \Rightarrow E; A \Rightarrow M)$ is *regular* if the core-anchor has constant rank. In this case $(W \Rightarrow E; A \Rightarrow M)$ is the direct sum of a *type* 1 and a *type* 0

VB-algebroid, up to isomorphisms. So type 1 and type 0 VB-algebroids are the building blocks of regular VB-algebroids. In this subsection we discuss linear deformation cohomologies of type 1 VB-algebroids.

Let $(W \Rightarrow E; A \Rightarrow M)$ be a type 1 VB-algebroid, and denote by $q : E \to M$ the projection. Gracia-Saz and Mehta [2010] show that $(W \Rightarrow E; A \Rightarrow M)$ is canonically isomorphic to the VB-algebroid $(q^!A \Rightarrow E; A \Rightarrow M)$. Here $q^!A \Rightarrow E$ is the *pull-back Lie algebroid*. Recall that its total space $q^!A$ is the fibered product $q^!A := TE \,{}_{dq}\!\times_{\rho} A$. Hence, sections of $q^!A \to E$ are pairs (X, α), where X is a vector field on E and α is a section of the pull-back bundle $q^*A \to E$, with the additional property that $dq(X_e) = \rho(\alpha_{q(e)})$ for all $e \in E$. Then there exists a unique Lie algebroid structure $q^!A \Rightarrow E$ such that the anchor $q^!A \to TE$ is the projection $(X, \alpha) \mapsto X$, and the Lie bracket is given by

$$[(X, q^*\alpha), (Y, q^*\beta)] = ([X, Y], q^*[\alpha, \beta]),$$

on sections of the special form $(X, q^*\alpha), (X, q^*\beta)$, with $\alpha, \beta \in \Gamma(A)$. Finally, $(q^!A \Rightarrow E; A \Rightarrow M)$ is a VB-algebroid, and every VB-algebroid of type 1 arises in this way (up to isomorphisms).

As $E \to M$ is a vector bundle, it has contractible fibers. So, according to [Sparano and Vitagliano 2018], $q^!A \Rightarrow E$ and $A \Rightarrow M$ share the same deformation cohomology. As an immediate consequence we get that the canonical map $C^\bullet_{\mathrm{def,lin}}(q^!A) \to C^\bullet_{\mathrm{def}}(A)$ induces an injection in cohomology. We want to show that it is a quasi-isomorphism. To do this it is enough to prove that the kernel $\mathfrak{End}(q^!A[1]_E)^\bullet$ of $C^\bullet_{\mathrm{def,lin}}(q^!A) \to C^\bullet_{\mathrm{def}}(A)$ is acyclic. We use graded geometry again. So, consider the pull-back diagram

$$\begin{array}{ccc} q^!A & \longrightarrow & TE \\ \downarrow & & \downarrow{\scriptstyle dq} \\ A & \xrightarrow{\rho} & TM \end{array}$$

All vertices are vector bundles, and shifting by one the degree in their fibers, we get a pull-back diagram of DG-manifolds:

$$\begin{array}{ccc} q^!A[1]_E & \longrightarrow & T[1]E \\ {\scriptstyle \tilde{q}}\downarrow & & \downarrow{\scriptstyle dq} \\ A[1] & \xrightarrow{\rho} & T[1]M \end{array}$$

This shows, among other things, that there is a canonical isomorphism

$$\mathfrak{End}(q^!A[1]_E)^\bullet = C^\bullet(A) \otimes_{\Omega^\bullet(M)} \mathfrak{End}(T[1]E)^\bullet$$

of DG-modules. From Proposition 1.4.5, exact sequence (1-16), and the fact that the deformation cohomology of $TM \Rightarrow M$ is trivial, $\mathfrak{End}(T[1]E)^\bullet$ is acyclic. Actually

there is a canonical contracting homotopy $h' : \mathfrak{End}(T[1]E)^\bullet \to \mathfrak{End}(T[1]E)^{\bullet-1}$. Indeed, there is a canonical contracting homotopy $H : \mathfrak{X}(T[1]E)^\bullet \to \mathfrak{X}(T[1]E)^{\bullet-1}$ restricting to both $\mathfrak{X}_{\mathrm{lin}}(T[1]E)^\bullet$ and $\mathfrak{End}(T[1]E)^\bullet$ (see, e.g., [Vitagliano 2014; Sparano and Vitagliano 2018], for a definition of H). Then, h' is simply the restriction of H, and it is graded $C^\bullet(A)$-linear. Finally, we define a contracting homotopy

$$h : \mathfrak{End}(q^!A[1]_E)^\bullet \to \mathfrak{End}(q^!A[1]_E)^{\bullet-1}$$

by putting $h(\omega \otimes \Phi) := (-)^\omega \omega \otimes h'(\Phi)$ for all $\omega \in C^\bullet(A)$, and all $\Phi \in \mathfrak{End}(T[1]E)^\bullet$. Summarizing, we have proved:

Proposition 2.6.1. *Let $(W \Rightarrow E; A \Rightarrow M)$ be a type 1 VB-algebroid. Then the canonical surjection $C^\bullet_{\mathrm{def,lin}}(W) \to C^\bullet_{\mathrm{def}}(A)$ is a quasi-isomorphism. In particular, $H^\bullet_{\mathrm{def,lin}}(W) = H^\bullet_{\mathrm{def}}(A)$.*

In essence, deforming a type 1 VB-algebroid is the same as deforming its base Lie algebroid.

Appendix: The homogeneity structure of a vector bundle

Here, for the reader's convenience, we recall the well-known concepts of homogeneity structure of a vector bundle and of linear multivectors on its total space. We make no claim of originality: these ideas appeared (probably for the first time) in [Grabowski and Rotkiewicz 2009], [Iglesias-Ponte et al. 2012] and [Bursztyn and Cabrera 2012], respectively. In the last two references, the reader can also find the proof of (a version of) Proposition A.0.3. With respect to those references, we will offer just a slightly different point of view, in order to make the presentation consistent. Notations and conventions in this appendix are used throughout the paper, sometimes without further comments.

Let $E \to M$ be a vector bundle. The monoid $\mathbb{R}_{\geq 0}$ of nonnegative real numbers acts on E by homotheties $h_\lambda : E \to E$ (fiber-wise scalar multiplication). The action $h : \mathbb{R}_{\geq 0} \times E \to E$, $e \mapsto h_\lambda(e)$, is called the *homogeneity structure* of E. The homogeneity structure (together with the smooth structure) fully characterizes the vector bundle structure [Grabowski and Rotkiewicz 2009]. In particular, it determines the addition. This implies that *every notion that involves the linear structure of E can be expressed in terms of h only*: for example, a smooth map between the total spaces of two vector bundles is a bundle map if and only if it commutes with the homogeneity structures.

The homogeneity structure isolates a distinguished subspace in the algebra $\mathfrak{X}^\bullet(E)$ of multivectors on the total space E of the vector bundle.

Definition A.0.1. A multivector $X \in \mathfrak{X}^\bullet(E)$ is (*homogeneous*) *of weight* q if and only if

$$h_\lambda^* X = \lambda^q X \tag{A-1}$$

for all $\lambda > 0$. The space of k-vector fields of weight q on E will be denoted $\mathfrak{X}^k_q(E)$. We denote simply by $C^\infty_q(E) := \mathfrak{X}^0_q(E)$ the space of functions of weight q and by $\mathfrak{X}_q(E) := \mathfrak{X}^1_q(E)$ the space of vector fields of weight q.

Clearly, for $q \geq 0$, weight q functions coincide with functions on E that are fiber-wise polynomial of degree q, while for $q < 0$ there are no nonzero functions of weight q. In particular, weight-zero functions are fiber-wise constant functions, i.e., pull-backs of functions on the base M. We refer to them as *core functions* and we denote $C^\infty_{\mathrm{core}}(E) := C^\infty_0(E)$.

The functorial properties of the pull-back imply that the grading defined by the weight is natural with respect to all the usual operations on functions and (multi)vector fields. From this remark, we easily see that *there are no nonzero k-vector fields of weight less than $-k$*.

Definition A.0.2. A function on E is *linear* if it is of weight 1. More generally, a k-vector field is *linear* if it is of weight $1-k$. We denote by $C^\infty_{\mathrm{lin}}(E)$, $\mathfrak{X}_{\mathrm{lin}}(E)$ and $\mathfrak{X}^\bullet_{\mathrm{lin}}(E)$ the spaces of linear functions, vector fields and multivectors, respectively.

Linear functions are precisely fiber-wise linear functions. The definition of linear multivectors may sound a little strange, but it is motivated (among other things) by the following proposition:

Proposition A.0.3. *Let $X \in \mathfrak{X}^k(E)$. The following conditions are equivalent*:

(1) *X is linear*;

(2) *X takes*
   (a) *k linear functions to a linear function*,
   (b) *$k-1$ linear functions and a core function to a core function*,
   (c) *$k-i$ linear functions and i core functions to* 0, *for every $i \geq 2$*;

(3) *If (x^i) are local coordinates on M and (u^α) are linear fiber coordinates on E, X is locally of the form*

$$X = X^{\alpha_1\cdots\alpha_{k-1}i}(x)\frac{\partial}{\partial u^{\alpha_1}}\wedge\cdots\wedge\frac{\partial}{\partial u^{\alpha_{k-1}}}\wedge\frac{\partial}{\partial x^i} + X^{\alpha_1\cdots\alpha_k}_\beta(x)u^\beta\frac{\partial}{\partial u^{\alpha_1}}\wedge\cdots\wedge\frac{\partial}{\partial u^{\alpha_k}}. \tag{A-2}$$

Acknowledgments

We thank the referee for reading carefully our manuscript, and for several suggestions that improved the presentation considerably. The authors are members of GNSAGA of INdAM.

References

[Arias Abad and Crainic 2012] C. Arias Abad and M. Crainic, "Representations up to homotopy of Lie algebroids", *J. Reine Angew. Math.* **663** (2012), 91–126. MR Zbl

[Behrend and Xu 2011] K. Behrend and P. Xu, "Differentiable stacks and gerbes", *J. Symplectic Geom.* **9**:3 (2011), 285–341. MR Zbl

[Bursztyn and Cabrera 2012] H. Bursztyn and A. Cabrera, "Multiplicative forms at the infinitesimal level", *Math. Ann.* **353**:3 (2012), 663–705. MR Zbl

[Cabrera and Drummond 2017] A. Cabrera and T. Drummond, "Van Est isomorphism for homogeneous cochains", *Pacific J. Math.* **287**:2 (2017), 297–336. MR Zbl

[Crainic and Moerdijk 2008] M. Crainic and I. Moerdijk, "Deformations of Lie brackets: cohomological aspects", *J. Eur. Math. Soc.* **10**:4 (2008), 1037–1059. MR Zbl

[Crainic et al. 2015] M. Crainic, J. N. Mestre, and I. Struchiner, "Deformations of Lie groupoids", preprint, 2015. arXiv

[Esposito et al. 2016] C. Esposito, A. G. Tortorella, and L. Vitagliano, "Infinitesimal automorphisms of VB-groupoids and algebroids", preprint, 2016. arXiv

[Grabowska et al. 2003] K. Grabowska, J. Grabowski, and P. Urbański, "Lie brackets on affine bundles", *Ann. Global Anal. Geom.* **24**:2 (2003), 101–130. MR Zbl

[Grabowski and Rotkiewicz 2009] J. Grabowski and M. Rotkiewicz, "Higher vector bundles and multi-graded symplectic manifolds", *J. Geom. Phys.* **59**:9 (2009), 1285–1305. MR Zbl

[Gracia-Saz and Mehta 2010] A. Gracia-Saz and R. A. Mehta, "Lie algebroid structures on double vector bundles and representation theory of Lie algebroids", *Adv. Math.* **223**:4 (2010), 1236–1275. MR Zbl

[del Hoyo and Ortiz 2016] M. del Hoyo and C. Ortiz, "Morita equivalences of vector bundles", preprint, 2016. arXiv

[Iglesias-Ponte et al. 2012] D. Iglesias-Ponte, C. Laurent-Gengoux, and P. Xu, "Universal lifting theorem and quasi-Poisson groupoids", *J. Eur. Math. Soc.* **14**:3 (2012), 681–731. MR Zbl

[La Pastina and Vitagliano 2019] P. P. La Pastina and L. Vitagliano, "Deformations of vector bundles over Lie groupoids", preprint, 2019. arXiv

[Mackenzie 1998a] K. C. H. Mackenzie, "Double Lie algebroids and the double of a Lie bialgebroid", preprint, 1998. arXiv

[Mackenzie 1998b] K. C. H. Mackenzie, "Drinfel'd doubles and Ehresmann doubles for Lie algebroids and Lie bialgebroids", *Electron. Res. Announc. Amer. Math. Soc.* **4** (1998), 74–87. MR Zbl

[Mackenzie 2005] K. C. H. Mackenzie, *General theory of Lie groupoids and Lie algebroids*, London Math. Soc. Lecture Note Series **213**, Cambridge Univ. Press, 2005. MR Zbl

[Mehta 2006] R. A. Mehta, *Supergroupoids, double structures, and equivariant cohomology*, Ph.D. thesis, University of California, Berkeley, 2006, available at https://search.proquest.com/docview/305363751.

[Nijenhuis and Richardson 1966] A. Nijenhuis and R. W. Richardson, Jr., "Cohomology and deformations in graded Lie algebras", *Bull. Amer. Math. Soc.* **72** (1966), 1–29. MR Zbl

[Nijenhuis and Richardson 1967a] A. Nijenhuis and R. W. Richardson, Jr., "Deformations of homomorphisms of Lie groups and Lie algebras", *Bull. Amer. Math. Soc.* **73** (1967), 175–179. MR Zbl

[Nijenhuis and Richardson 1967b] A. Nijenhuis and R. W. Richardson, Jr., "Deformations of Lie algebra structures", *J. Math. Mech.* **17** (1967), 89–105. MR Zbl

[Pradines 1967] J. Pradines, "Théorie de Lie pour les groupoïdes différentiables: calcul différentiel dans la catégorie des groupoïdes infinitésimaux", *C. R. Acad. Sci. Paris Sér. A-B* **264** (1967), 245–248. MR Zbl

[Sparano and Vitagliano 2018] G. Sparano and L. Vitagliano, "Deformation cohomology of Lie algebroids and Morita equivalence", *C. R. Math. Acad. Sci. Paris* **356**:4 (2018), 376–381. MR Zbl

[Vaintrob 1997] A. Y. Vaintrob, "Lie algebroids and homological vector fields", *Uspekhi Mat. Nauk* **52**:2(314) (1997), 161–162. In Russian; translated in *Russian Math. Surv.* **52**:2 (1997), 428–429. MR Zbl

[Vitagliano 2014] L. Vitagliano, "On the strong homotopy Lie–Rinehart algebra of a foliation", *Commun. Contemp. Math.* **16**:6 (2014), art. id. 1450007. MR Zbl

[Vitagliano 2016] L. Vitagliano, "Vector bundle valued differential forms on $\mathbb{N}Q$-manifolds", *Pacific J. Math.* **283**:2 (2016), 449–482. MR Zbl

[Voronov 2012] T. T. Voronov, "Q-manifolds and Mackenzie theory", *Comm. Math. Phys.* **315**:2 (2012), 279–310. MR Zbl

Received June 4, 2018. Revised March 1, 2019.

PIER PAOLO LA PASTINA
DEPARTMENT OF MATHEMATICS
SAPIENZA UNIVERSITY OF ROME
ROME
ITALY
lapastina@mat.uniroma1.it

LUCA VITAGLIANO
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SALERNO
FISCIANO
ITALY
lvitagliano@unisa.it

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.299

A MOD-p ARTIN–TATE CONJECTURE, AND GENERALIZING THE HERBRAND–RIBET THEOREM

DIPENDRA PRASAD

We propose conjectures about the integrality properties of the values at $s = 0$ of certain abelian L-functions of $\mathbb{Q}$ and totally real number fields. We also propose a conjecture which generalizes the theorems of Herbrand and Ribet for values at $s = 0$ of totally odd Artin L-functions of totally real number fields. Various calculations, some of which are familiar to experts, are made to provide examples.

1. Introduction

Following the natural instinct that when a group operates on a number field then every term in the class number formula should factorize "compatibly" according to the representation theory (both complex and modular) of the group, we are led — in the spirit of Herbrand and Ribet's theorem on the p-component of the class number of $\mathbb{Q}(\zeta_p)$ — to some natural questions about the p-part of the class group of any CM Galois extension E of $\mathbb{Q}$ as a module for $\mathrm{Gal}(E/\mathbb{Q})$. The compatible factorization of the class number formula is at the basis of *Stark's conjecture*, where one is mostly interested in factorizing the regulator term — whereas for us in this paper, we put ourselves in a situation where the regulator term can be ignored, and it is the factorization of the class number that we seek. All this is presumably part of various "equivariant" conjectures in arithmetic-geometry, such as the "equivariant Tamagawa number conjecture", but the literature does not seem to address this question in any precise way. In trying to formulate these questions, we are naturally led to consider $L(0, \rho)$, for ρ an Artin representation, in situations where this is known to be nonzero and algebraic, and it is important for us to understand if this is $\mathfrak{p}$-integral for a prime $\mathfrak{p}$ of the ring of algebraic integers $\overline{\mathbb{Z}}$ in $\mathbb{C}$, which we call a *mod-p Artin–Tate conjecture*. As an attentive reader will notice, the most innocuous term in the class number formula, the number of roots of unity, plays an important role for us — it, being the only term in the denominator, is responsible for all the poles!

MSC2010: 11F33, 11R23.

Keywords: class groups, class number formula, Iwasawa theory, main conjecture, L-values, algebraicity of L-values, Herbrand–Ribet theorem.

Let F be a number field contained in $\mathbb{C}$ with $\overline{\mathbb{Q}}$ its algebraic closure in $\mathbb{C}$. Let $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_n(\mathbb{C})$ be an irreducible Galois representation with $L(s, \rho)$ its associated Artin L-function. According to a famous conjecture of Artin, $L(s, \rho)$ has an analytic continuation to an entire function on $\mathbb{C}$ unless ρ is the trivial representation, in which case it has a unique pole at $s = 1$ which is simple.

More generally, let M be an irreducible motive over $\mathbb{Q}$ with $L(s, M)$ its associated L-function. According to Tate, $L(s, M)$ has an analytic continuation to an entire function on $\mathbb{C}$ unless M is a twisted Tate motive $\mathbb{Q}(j)$ with $\mathbb{Q}(1)$ the motive associated to $\mathbb{G}_m$. For the motive $\mathbb{Q} = \mathbb{Q}(0)$, $L(s, \mathbb{Q}) = \zeta_{\mathbb{Q}}(s)$, the usual Riemann zeta function, which has a unique pole at $s = 1$ which is simple.

This paper will deal with certain Artin representations $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_n(\mathbb{C})$ for which we will know a priori that $L(0, \rho)$ is a nonzero algebraic number (in particular, F will be totally real). It is then an important question to understand the nature of the algebraic number $L(0, \rho)$: to know if it is an algebraic integer, but if not, what are its possible denominators. We think of the possible denominators in $L(0, \rho)$, as existence of poles for $L(0, \rho)$, at the corresponding prime ideals of $\overline{\mathbb{Z}}$. It is thus analogous to the conjectures of Artin and Tate, both in its aim — and as we will see — in its formulation. Since we have chosen to understand L-values at 0 instead of 1, which is where Artin and Tate conjectures are formulated, there is an "ugly" twist by ω_p — the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on the p-th roots of unity — throughout the paper, giving a natural character $\omega_p : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to (\mathbb{Z}/p)^\times$, also a character of $\mathrm{Gal}(\overline{\mathbb{Q}}/L)$ for L any algebraic extension of $\mathbb{Q}$, as well as a character of $\mathrm{Gal}(L/\mathbb{Q})$ if L is a Galois extension of $\mathbb{Q}$ containing p-th roots of unity; if there are no nontrivial p-th roots of unity in L, we will define ω_p to be the trivial character of $\mathrm{Gal}(L/\mathbb{Q})$.

We now fix some notation. We will fix an isomorphism of $\overline{\mathbb{Q}}_p$ with $\mathbb{C}$ where $\overline{\mathbb{Q}}_p$ is a fixed algebraic closure of $\mathbb{Q}_p$, the field of p-adic numbers. This allows one to define $\mathfrak{p}$, a prime ideal in $\overline{\mathbb{Z}}$, the integral closure of $\mathbb{Z}$ in $\mathbb{C}$, over the prime ideal generated by p in $\mathbb{Z}$. The prime p will always be an odd prime in this paper.

All the finite-dimensional representations of finite groups in this paper will take values in $\mathrm{GL}_n(\overline{\mathbb{Q}}_p)$, and therefore in $\mathrm{GL}_n(\mathbb{C})$, as well as $\mathrm{GL}_n(\overline{\mathbb{Z}}_p)$. It thus makes sense to talk of "reduction modulo $\mathfrak{p}$" of (complex) representations of finite groups. These reduced representations are well defined up to semisimplification on vector spaces over $\overline{\mathbb{F}}_p$ (theorem of Brauer and Nesbitt); we denote the reduction modulo $\mathfrak{p}$ of representations as $\rho \to \bar{\rho}$.

If F is a finite Galois extension of $\mathbb{Q}$ with Galois group G, then it is well known that the zeta function $\zeta_F(s)$ can be factorized as

$$\zeta_F(s) = \prod_{\rho} L(s, \rho)^{\dim \rho},$$

where ρ ranges over all the irreducible complex representations of G, and $L(s, \rho)$ denotes the Artin L-function associated to ρ.

According to the class number formula, we have the power series expansion of $\zeta_F(s)$ at $s = 0$ as

$$\zeta_F(s) = -\frac{hR}{w} s^{r_1+r_2-1} + \text{higher-order terms},$$

where r_1, r_2, h, R, w are the standard invariants associated to F: r_1 the number of real embeddings, r_2 the number of pairs of complex conjugate embeddings which are not real, h the class number of F, R the regulator, and w the number of roots of unity in F.

This paper considers ζ_E/ζ_F where E is a CM field with F its totally real subfield, in which case $r_1 + r_2$ is the same for E as for F, and the regulators of E and F too are the same except for a possible power of 2. Therefore, for c the complex conjugation on $\mathbb{C}$,

$$(\zeta_E/\zeta_F)(0) = \prod_{\rho(c)=-1} L(0, \rho)^{\dim \rho} = \frac{h_E/h_F}{w_E/w_F},$$

where each L-value $L(0, \rho)$ in the above expression is a nonzero algebraic number by a theorem of Klingen and Siegel.

In this identity, observe that L-functions are associated to $\mathbb{C}$-representations of $\mathrm{Gal}(E/\mathbb{Q})$, whereas the class groups of E and F are finite Galois modules. Modulo some details, we basically assert that for each odd prime p, each irreducible odd $\mathbb{C}$-representation ρ of $\mathrm{Gal}(E/\mathbb{Q})$ contributes a certain number of copies (depending on $\mathfrak{p}$-adic valuation of $L(0, \rho)$) of $\bar{\rho}$ to the class group of E tensored with $\mathbb{F}_p$ modulo the class group of F tensored with $\mathbb{F}_p$ (up to semisimplification). This is exactly what happens for $E = \mathbb{Q}(\zeta_p)$ by the theorems of Herbrand and Ribet, which is the main motivating example for all that we do here, and this is what we will review next.

2. The Herbrand–Ribet theorem

In this section we recall the Herbrand–Ribet theorem from the point of view of this paper. We refer to the original work of Ribet [1976] and to [Washington 1982] for an exposition on the theorem together with a proof of Herbrand's theorem.

There are actually two a priori important aspects of the Herbrand–Ribet theorem dealing with the p-component of the class group for $\mathbb{Q}(\zeta_p)$. First, as the Galois group $\mathrm{Gal}(\mathbb{Q}(\zeta_p)/\mathbb{Q}) = (\mathbb{Z}/p)^\times$ is of order coprime to p, its action on the p-component of the class group is semisimple, and therefore, the p-component of the class group can be written as a direct sum of eigenspaces for the action of $(\mathbb{Z}/p)^\times$ on it. We do not consider this aspect of the Herbrand–Ribet theorem to be important, and simply consider semisimplification of representations of Galois groups on class groups to be a good-enough substitute.

The second — and more serious — aspect of the Herbrand–Ribet theorem is that among the characters of $\mathrm{Gal}(\mathbb{Q}(\zeta_p)/\mathbb{Q}) = (\mathbb{Z}/p)^\times$, only the odd characters, i.e., characters $\chi : (\mathbb{Z}/p)^\times \to \mathbb{Q}_p^\times$ with $\chi(-1) = -1$, present themselves — as it is only for these that there is any result about the χ-eigencomponent in the class group, and even among these, the Teichmüller character $\omega_p : (\mathbb{Z}/p)^\times \to \mathbb{Q}_p^\times$ plays a role different from other characters of $(\mathbb{Z}/p)^\times$. (Note that earlier we have used ω_p for the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on the p-th roots of unity, giving a natural character $\omega_p : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to (\mathbb{Z}/p)^\times$, as well as to its restriction to $\mathrm{Gal}(\overline{\mathbb{Q}}/L)$ for L any algebraic extension of $\mathbb{Q}$. Since $\mathrm{Gal}(\mathbb{Q}(\zeta_p)/\mathbb{Q})$ is canonically isomorphic to $(\mathbb{Z}/p)^\times$, the two roles that ω_p will play throughout the paper are actually the same.)

To elaborate on the role of "odd" characters in the Herbrand–Ribet theorem, observe that the class number formula

$$\zeta_F(s) = -\frac{hR}{w} s^{r_1+r_2-1} + \text{higher-order terms}$$

can be considered both for $F = \mathbb{Q}(\zeta_p)$ as well as its maximal real subfield $F^+ = \mathbb{Q}(\zeta_p)^+$. It is known that [Washington 1982, Proposition 4.16],

$$R/R^+ = 2^{(p-3)/2},$$

where R is the regulator for $\mathbb{Q}(\zeta_p)$ and R^+ is the regulator for $\mathbb{Q}(\zeta_p)^+$. We will similarly denote h and h^+ to be the order of the two class groups, with $h^- = h/h^+$, an integer.

Dividing the class number formula of $\mathbb{Q}(\zeta_p)$ by that of $\mathbb{Q}(\zeta_p)^+$, we find

$$\prod_{\chi \text{ an odd character of } (\mathbb{Z}/p)^\times} L(0, \chi) = \frac{1}{p} \cdot \frac{h}{h^+} \cdot 2^{(p-3)/2}, \tag{1}$$

the factor $1/p$ arising because there are $2p$ roots of unity in $\mathbb{Q}(\zeta_p)$ and only 2 in $\mathbb{Q}(\zeta_p)^+$.

It is known that for χ an odd character of $(\mathbb{Z}/p)^\times$, $L(0, \chi)$ is an algebraic number which is given in terms of the generalized Bernoulli number $B_{1,\chi}$ as

$$L(0, \chi) = -B_{1,\chi} = -\frac{1}{p} \sum_{a=1}^{a=p} a\chi(a).$$

It is easy to see that $pB_{1,\omega_p^{p-2}} \equiv (p-1) \bmod p$ since $a\omega_p^{p-2}(a)$ is the trivial character of $(\mathbb{Z}/p)^\times$ whereas for all the other characters of $(\mathbb{Z}/p)^\times$, $L(0, \chi)$ is not only an algebraic number but is p-adic integral (Schur orthogonality!); all this is clear by looking at the expression

$$L(0, \chi) = -B_{1,\chi} = -\frac{1}{p} \sum_{a=1}^{a=p} a\chi(a).$$

Rewrite (1) up to p-adic units as

$$\prod_{\substack{\chi \text{ an odd character of } (\mathbb{Z}/p)^\times \\ \chi \neq \omega_p^{p-2} = \omega_p^{-1}}} L(0, \chi) = \frac{h}{h^+},$$

where we note that both sides of the equality are p-adic integral elements; in fact, since all characters $\chi : (\mathbb{Z}/p)^\times \to \overline{\mathbb{Q}}_p^\times$ take values in $\mathbb{Z}_p$, for $\chi \neq \omega_p^{-1}$, $L(0, \chi) \in \mathbb{Z}_p$. This, when interpreted — just an interpretation in the optics of this paper without any suggestions for proof in either direction! — for each χ component on the two sides of this equality, amounts to the theorem of Herbrand and Ribet which asserts that p divides $L(0, \chi) = -B_{1,\chi}$ for χ an odd character of $(\mathbb{Z}/p)^\times$, which is not ω_p^{p-2}, if and only if the corresponding χ^{-1}-eigencomponent of the class group of $\mathbb{Q}(\zeta_p)$ is nontrivial (note the χ^{-1}, and not χ!). Furthermore, the character ω_p does not appear in the p-class group of $\mathbb{Q}(\zeta_p)$. It can happen that $L(0, \chi)$ is divisible by higher powers of p than 1, and one expects — this is not proven yet! — that in such cases, the corresponding χ^{-1}-eigencomponent of the class group of $\mathbb{Q}(\zeta_p)$ is $\mathbb{Z}/p^{(\mathrm{val}_p L(0,\chi))}$, and in particular, it still has p-rank 1. (By [Mazur and Wiles 1984], the χ^{-1}-eigencomponent of the class group of $\mathbb{Q}(\zeta_p)$ is of order $p^{(\mathrm{val}_p L(0,\chi))}$.)

The work of Ribet was to prove that if $p \mid B_{1,\chi}$, then the χ^{-1}-eigencomponent of the class group of $\mathbb{Q}(\zeta_p)$ is nontrivial by constructing an unramified extension of $\mathbb{Q}(\zeta_p)$ by using a congruence between a holomorphic cusp form and an Eisenstein series on $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$.

To be able to use the class number formula in other situations, we will need to have the integrality of $L(0, \chi)$ for χ a character associated to the Galois group of a number field, or even of $L(0, \rho)$ for general irreducible representations ρ of the Galois group of a number field, in more situations that we call a mod-p Artin–Tate conjecture.

Let E be a CM number field which we assume is Galois over $\mathbb{Q}$. Assume that E contains p^n-th roots of unity but no p^{n+1}-th root of unity. Let F be the totally real subfield of E with $[E : F] = 2$. Let $G = \mathrm{Gal}(E/Q)$ with $-1 \in G$, the complex conjugation in G.

We have

$$\zeta_E(s) = \prod_\rho L_\mathbb{Q}(s, \rho)^{\dim \rho},$$

$$\zeta_F(s) = \prod_{\rho(-1)=1} L_\mathbb{Q}(s, \rho)^{\dim \rho},$$

$$(\zeta_E/\zeta_F)(s) = \prod_{\rho(-1)=-1} L_\mathbb{Q}(s, \rho)^{\dim \rho},$$

where all the products above are over irreducible representations ρ of $G = \mathrm{Gal}(E/\mathbb{Q})$ with values in $\mathrm{GL}_d(\overline{\mathbb{Q}}_p)$.

By the class number formula,

$$h^-(E)/p^n = \prod_{\rho(-1)=-1} L_{\mathbb{Q}}(0,\rho)^{\dim \rho}. \tag{2}$$

By Corollary 6 below, there is an $a \in \mathbb{Z}_p^\times$ with

$$a/p^n = \prod_{\chi(-1)=-1} L_{\mathbb{Q}}(0,\chi), \tag{3}$$

where χ are all the characters of $\mathrm{Gal}(\mathbb{Q}(\zeta_{p^n})/\mathbb{Q}) = (\mathbb{Z}/p^n)^\times$ for which $\bar{\chi} = \omega_p^{-1}$.

Dividing (2) by (3), we have up to p-adic units

$$h^-(E) = \prod_{\substack{\rho(-1)=-1,\\ \rho\neq\chi}} L_{\mathbb{Q}}(0,\rho)^{\dim \rho} \tag{4}$$

where the product on the right is taken over irreducible representations ρ of $G = \mathrm{Gal}(E/\mathbb{Q})$ for which $\rho(-1) = -1$, and which are not cyclotomic characters of the form $\chi : \mathrm{Gal}(\mathbb{Q}(\zeta_{p^n})/\mathbb{Q}) = (\mathbb{Z}/p^n)^\times \to \mathbb{Q}_p^\times$ with $\bar{\chi} = \omega_p^{-1}$.

It is known that $L(0,\rho) \in \overline{\mathbb{Q}}^\times$ for $\rho(-1) = -1$. This is a simple consequence of a theorem due to Klingen and Siegel that partial zeta functions of a totally real number field take rational values at all nonpositive integers [Tate 1984]. (Note that to prove $L(0,\rho) \in \overline{\mathbb{Q}}^\times$ for $\rho(-1) = -1$, it suffices by Brauer to prove it for abelian CM extensions by a lemma of Serre [Tate 1984, Chapter III, Lemma 1.3].)

The left-hand side of (4) is integral, and we would like to suggest the same for each term on the right-hand side of (4).

The following conjecture about $L(0,\rho)$ extends the known integrality properties of $L(0,\chi) = -B_{1,\chi} = -\frac{1}{p}\sum_{a=1}^{a=p} a\chi(a)$, encountered and used earlier. The formulation of the conjecture also assumes known integrality properties about $L(0,\chi)$ for $\chi : \mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) = (\mathbb{Z}/n)^\times \to \mathbb{C}^\times$ discussed in the last section of this paper.

Conjecture 1 (mod-p analogue of the Artin–Tate conjecture). *Let* $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_n(\mathbb{C})$ *be an irreducible representation of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ *with* $\rho(c) = -1$ *where* c *is a complex conjugation in* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. *Then unless* ρ *is a one-dimensional representation factoring through* $\mathrm{Gal}(\mathbb{Q}(\zeta_{p^n})/\mathbb{Q})$ (*for some prime* p) *with* $\bar{\rho}$ *the reduction of* ρ *modulo* $\mathfrak{p}$ (*the maximal ideal in* $\overline{\mathbb{Z}}_p$) *being* $\bar{\rho} = \omega_p^{-1}$, $L(0,\rho) \in \overline{\mathbb{Q}}$ *is integral outside* 2, *i.e.*, $L(0,\rho) \in \overline{\mathbb{Z}}\left[\frac{1}{2}\right]$.

We next recall the following theorem of Deligne and Ribet [1980], which could be considered as a weaker version of Conjecture 1.

Theorem. *Let* k *be a totally real number field, and let* $\chi : \mathrm{Gal}(\overline{\mathbb{Q}}/k) \to \overline{\mathbb{Q}}^\times$ *be a character of finite order with* $\chi(c) = -1$ *where* c *is a complex conjugation*

in $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$. *Let* w *be the order of the group of roots of unity in* E, *the smallest extension of* k *such that* χ *is trivial when restricted to* $\mathrm{Gal}(\overline{\mathbb{Q}}/E)$. *Then*

$$wL(0, \chi) \in \overline{\mathbb{Z}}.$$

In fact Conjecture 1 can be used to make precise the above theorem of Deligne–Ribet as follows; the simple argument using the fact that the Artin L-function is invariant under induction from $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$ to $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ will be left to the reader.

Conjecture 2. *Let* k *be a totally real number field, and* $\chi : \mathbb{A}_k^\times / k^\times \to \overline{\mathbb{Z}}_p^\times$ *a finite-order character, with* $\chi(c) = -1$ *where* c *is a complex conjugation in* $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$. *Then if* $L(0, \chi) \notin \overline{\mathbb{Z}}_p$,

(1) $\chi \mod p$ *is* ω_p^{-1} *and*

(2) χ *is a character of* $\mathbb{A}_k^\times / k^\times$ *associated to a character of the Galois group* $\mathrm{Gal}(k(\zeta_q)/k)$ *for some* q *which is a power of* p.

Remark. In the examples that I know, which are for characters $\chi : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \overline{\mathbb{Q}}_p^\times$ with $\chi = \omega_p^{-1} \pmod p$, if $L(0, \chi)$ has a (mod-p) pole, the pole is of order 1; more precisely, if $L = \mathbb{Q}_p[\chi(\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p))]$ is the subfield of $\overline{\mathbb{Q}}_p$ generated by the image under χ of the decomposition group at p, then $L(0, \chi)$ is the inverse of a uniformizer of this field L. It would be nice to know if this is the case for characters χ of $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$ for k arbitrary. This would be in the spirit of Artin's classical conjecture where the only possible poles of $L(1, \rho)$, for ρ an irreducible representation of $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$, are simple.

3. Proposed generalization of Herbrand–Ribet for CM number fields

The Herbrand–Ribet theorem is about the relationship of L-values $L(0, \chi)$ with the χ^{-1}-eigencomponent of the class group of $\mathbb{Q}(\zeta_p)$. In the last section, we proposed a precise conjecture about integrality properties for the L values $L(0, \rho)$. In this section, we now propose their relationship to class groups.

We begin by introducing some notation involved in constructing in a functorial way an elementary abelian p-group $\bar{A}[p]$ out of a finite abelian group A with

(1) $p \cdot \bar{A}[p] = 0$ and

(2) the cardinality of $\bar{A}[p]$ equals the cardinality of the p-Sylow subgroup of A.

We define $\bar{A}[p]$ to be the direct sum of the p-groups $p^i A / p^{i+1} A$ for $i \geq 0$. If A is a G-module, then naturally, $\bar{A}[p]$ too is a G-module. If A is a G-module, then we let $\bar{A}[p]^{\mathrm{ss}}$ be the semisimplification of the corresponding G-module $\bar{A}[p]$ over $\mathbb{F}_p$.

Since according to the theorem of Klingen and Siegel, the value $L(0, \rho)$ for an odd representation ρ of $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$, where k is a totally real number field, belongs to

the algebraic number field generated by the character values of ρ, and since we are trying to equate powers of p appearing on the two sides of the class number formula, it will be important to consider only those representations $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ which actually take values in $\mathbb{Q}_p^{\mathrm{unr}}$, the maximal unramified extension of $\mathbb{Q}_p$. Observe that the Brauer group of $\mathbb{Q}_p^{\mathrm{unr}}$ is trivial, and thus, an irreducible representation of a finite group is defined over $\mathbb{Q}_p^{\mathrm{unr}}$ if and only if its character is defined over $\mathbb{Q}_p^{\mathrm{unr}}$. If an irreducible representation π of a finite group is defined over $\overline{\mathbb{Q}}_p$, one can take the sum of the Galois conjugates π^σ of π for $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p^{\mathrm{unr}})$, to construct canonically an irreducible representation, say $\langle \pi \rangle$ over $\mathbb{Q}_p^{\mathrm{unr}}$. The representation π can be reduced modulo $\mathfrak{p}$ and the representation $\langle \pi \rangle$ modulo p, and the semisimplification of these reductions are related by

$$\overline{\langle \pi \rangle}^{\mathrm{ss}} \cong d\bar{\pi}^{\mathrm{ss}},$$

where d is the number of distinct Galois conjugates of π under $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p^{\mathrm{unr}})$.

Let E be a Galois CM extension of a totally real number field k with F the totally real subfield of E with $[E : F] = 2$, and $G = \mathrm{Gal}(E/k)$. Let c denote the element of order 2 in the Galois group of E over F.

Let H_E and H_F denote the class groups of E and F, respectively. Observe that the kernel of the natural map from H_F to H_E is a 2-group. (This follows from using the norm mapping from H_E to H_F.) Therefore, since we are interested in p-primary components for only odd primes p, H_F can be considered to be a subgroup of H_E, and the quotient H_E/H_F becomes a G-module of order h_E/h_F.

The following conjecture on the structure of the *minus-part* of the class group of E (as a module for the Galois group G) is arrived at by considering the p-adic valuations of the two sides of the class number formula

$$(\zeta_E/\zeta_F)(0) = \prod_{\rho(c)=-1} L(0, \rho)^{\dim \rho} = \prod_{\rho(c)=-1} L(0, \langle \rho \rangle)^{\dim \rho} = \frac{h_E}{h_F} \frac{1}{w_E},$$

with E, F, k as above, and the first product taken over all irreducible representations ρ of $G = \mathrm{Gal}(E/k)$ with values in $\mathrm{GL}_n(\overline{\mathbb{Q}}_p)$, whereas the second one is over all irreducible representations $\langle \rho \rangle$ of $G = \mathrm{Gal}(E/k)$ with values in $\mathrm{GL}_n(\mathbb{Q}_p^{\mathrm{unr}})$. Since we are formulating the conjecture below based on equality of (p-adic valuations of) numbers in the class number formula, it is not sensitive to the subtlety discussed earlier about χ-eigencomponents in the class group of $\mathbb{Q}(\mu_p)$ being cyclic or not; all we care about is their order.

Conjecture 3. *Let E be a CM, Galois extension of a totally real number field k, with F the totally real subfield of E, and $c \in \mathrm{Gal}(E/F)$, the nontrivial element of the Galois group. Let $\langle \rho \rangle : \mathrm{Gal}(E/k) \to \mathrm{GL}_{dn}(\mathbb{Q}_p^{\mathrm{unr}})$ be an irreducible, odd (i.e., $\rho(c) = -1$) representation of $\mathrm{Gal}(E/k)$ associated to an irreducible representation $\rho : \mathrm{Gal}(E/k) \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ as above, with $\bar{\rho}$ the semisimplification of the reduction*

of $\rho \bmod \mathfrak{p}$ *for* p *an odd prime. Let* $\omega_p : \mathrm{Gal}(E/k) \to (\mathbb{Z}/p)^\times$ *be the action of* $\mathrm{Gal}(E/k)$ *on the* p*-th roots of unity in* E (*so* $\omega_p = 1$ *if* $\zeta_p \notin E$). *Then*

$$\overline{H_E/H_F}[p]^{\mathrm{ss}} = \sum_{\langle\rho\rangle} v_p(L(0, \langle\rho\rangle))\bar{\rho}^\vee,$$

an equality of representations of $\mathrm{Gal}(E/k)$, *except for the* ω_p*-component* ($\bar{\rho}^\vee$ *denotes the contragredient of* $\bar{\rho}$). *If* $\omega_p \neq 1$, *we make no assertion on the* ω_p*-component in* $\overline{H_E/H_F}[p]^{\mathrm{ss}}$, *but if* $\omega_p = 1$, *there is no* ω_p*-component inside* $\overline{H_E/H_F}[p]^{\mathrm{ss}}$.

Remark. For absolutely abelian fields, i.e., in the notation above, if E is an abelian extension of $\mathbb{Q}$, the conjecture above is known, and amounts to a conjecture of Gras [1977] which, for p not dividing the order of the Galois group, is proved by Mazur and Wiles [1984] as a consequence of their proof of the main conjecture, and for p dividing the order of the Galois group it is due to Solomon [1990].

4. Integrality of abelian L-values for $\mathbb{Q}$

The aim of this section is to prove certain results on integrality of $L(0, \chi)$ for χ an odd Dirichlet character of $\mathbb{Q}$ which are first examples of all the integrality conjectures made in this paper. Although these are all well known results, we have decided to give our proofs.

Lemma 4. *For integers* $m > 1$ *and* $n > 1$, *with* $(m, n) = 1$, *let* $\chi = \chi_1 \times \chi_2$ *be a primitive Dirichlet character on* $(\mathbb{Z}/mn\mathbb{Z})^\times = (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$ *with* $\chi(-1) = -1$. *Then*

$$L(0, \chi) = -B_{1,\chi} = -\frac{1}{mn}\sum_{a=1}^{mn} a\chi(a)$$

is an algebraic integer, *i.e.*, *belongs to* $\overline{\mathbb{Z}} \subset \overline{\mathbb{Q}}$.

Proof. Observe that $B_{1,\chi} = \frac{1}{mn}\sum_{a=1}^{mn} a\chi(a)$ has a possible fraction by mn, and that in this sum over $a \in \{1, 2, \ldots, mn\}$, if we instead sum over an arbitrary set A of integers which have these residues mod mn, then $\frac{1}{mn}\sum_{a\in A} a\chi(a)$ will differ from $B_{1,\chi}$ by an integral element (in $\overline{\mathbb{Z}}$). Since our aim is to prove that $B_{1,\chi}$ is integral, it suffices to prove that $\frac{1}{mn}\sum_{a\in A} a\chi(a)$ is integral for some set of representatives $A \subset \mathbb{Z}$ of residues mod mn.

For an integer $a \in \{1, 2, \ldots, m\}$, let $\bar{a}$ be an arbitrary integer whose reduction mod m is a, and whose reduction mod n is 1. Similarly, for an integer $b \in \{1, 2, \ldots, n\}$, let $\bar{b}$ be an arbitrary integer whose reduction mod n is b and whose reduction mod m is 1. Clearly, the set of integers $\bar{a} \cdot \bar{b}$ represents — exactly once — each residue class mod mn, and that $\bar{a} \cdot \bar{b}$ as an element in $\mathbb{Z}$ goes to the pair $(a, b) \in \mathbb{Z}/m \times \mathbb{Z}/n$. (It is important to note that $\bar{a} \cdot \bar{b}$ as an element in $\mathbb{Z}$ is *not* congruent

to $ab \bmod mn$, and therein lies a subtlety in the Chinese remainder theorem: there is no simple inverse to the natural isomorphism: $\mathbb{Z}/mn \to \mathbb{Z}/m \times \mathbb{Z}/n$.)

By definition of the character χ, $\chi(\bar{a} \cdot \bar{b}) = \chi_1(a)\chi_2(b)$. It follows that

$$\frac{1}{mn} \sum \bar{a}\bar{b}\chi(\bar{a} \cdot \bar{b}) - \left[\frac{1}{m} \sum_{a=1}^{m} a\chi_1(a)\right] \cdot \left[\frac{1}{n} \sum_{b=1}^{n} b\chi_2(b)\right] \in \overline{\mathbb{Z}}. \tag{5}$$

Since the character χ is odd, one of the characters, say χ_2, is even (and χ_1 is odd).

Observe that

$$B_{1,\chi_2} = \frac{1}{n} \sum_{b=1}^{n} b\chi_2(b) = \frac{1}{n} \sum_{b=1}^{n} (n-b)\chi_2(b).$$

It follows that

$$\frac{2}{n} \sum_{b=1}^{n} b\chi_2(b) = \sum_{b=1}^{n} \chi_2(b) = 0,$$

where the last sum is zero because the character χ_2 is assumed to be nontrivial.

Since

$$\frac{1}{mn} \sum \bar{a}\bar{b}\chi(\bar{a} \cdot \bar{b}) - \frac{1}{mn} \sum_{c=1}^{mn} c\chi(c) \in \overline{\mathbb{Z}},$$

by (5), it follows that

$$\frac{1}{mn} \sum_{c=1}^{mn} c\chi(c) \in \overline{\mathbb{Z}},$$

as desired. □

Lemma 5. *For p a prime, let χ be a primitive Dirichlet character on $(\mathbb{Z}/p^n\mathbb{Z})^\times$ with $\chi(-1) = -1$. Write $(\mathbb{Z}/p^n\mathbb{Z})^\times = (\mathbb{Z}/p\mathbb{Z})^\times \times (1 + p\mathbb{Z}/1 + p^n\mathbb{Z})$, and the character χ as $\chi_1 \times \chi_2$ with respect to this decomposition. Then*

$$L(0, \chi) = -B_{1,\chi} = -\frac{1}{p^n} \sum_{a=1}^{p^n} a\chi(a)$$

is an algebraic integer, i.e., belongs to $\overline{\mathbb{Z}} \subset \overline{\mathbb{Q}}$ if and only if $\chi_1 \neq \omega_p^{-1}$. Further, if $\chi_1 = \omega_p^{-1}$,

$$L(0, \chi) = -B_{1,\chi} = -\frac{1}{p^n} \sum_{a=1}^{p^n} a\chi(a)$$

is the inverse of a uniformizer in the local field $\mathbb{Q}_p(B_{1,\chi}) = \mathbb{Q}_p(\chi)$ which is a totally ramified cyclic extension of $\mathbb{Q}_p$ of degree equal to the order of χ_2.

Proof. Assuming that $\chi_1 \neq \omega_p^{-1}$, we prove that $B_{1,\chi}$ belongs to $\overline{\mathbb{Z}} \subset \overline{\mathbb{Q}}$.

By an argument similar to the one used in the previous lemma, it can be checked that

$$\frac{1}{p^n}\sum_{a=1}^{p^n} a\chi(a) - \left[\frac{1}{p}\sum_{a=1}^{p} a\chi_1(a)\right]\cdot\left[\frac{1}{p^{n-1}}\sum_{b=1}^{p^{n-1}}(1+bp)\chi_2(1+bp)\right] \in \overline{\mathbb{Z}}.$$

If $\chi_1 \neq \omega_p^{-1}$, $\frac{1}{p}\sum_{a=1}^{p} a\chi_1(a)$ is easily seen to be integral. To prove the lemma, it then suffices to prove that $\left[\frac{1}{p^{n-1}}\sum_{b=1}^{p^{n-1}}(1+bp)\chi_2(1+bp)\right]$ is integral.

Note the isomorphism of the additive group $\mathbb{Z}_p$ with the multiplicative group $1+p\mathbb{Z}_p$ by the map $n \to (1+p)^n \in 1+p\mathbb{Z}_p$. Let $\chi_2(1+p) = \alpha$ with $\alpha^{p^{n-1}} = 1$.

Then (the first and third equalities below are up to $\overline{\mathbb{Z}}$)

$$\begin{aligned}
\frac{1}{p^{n-1}}\sum_{b=1}^{p^{n-1}}(1+bp)\chi_2(1+bp) &= \frac{1}{p^{n-1}}\sum_{c=1}^{p^{n-1}}(1+p)^c\alpha^c \\
&= \frac{1}{p^{n-1}}\sum_{c=1}^{p^{n-1}}[\alpha(1+p)]^c \\
&= \frac{1}{p^{n-1}}\frac{1-[\alpha(1+p)]^{p^{n-1}}}{1-\alpha(1+p)} \\
&= \frac{1}{p^{n-1}}\frac{[1-(1+p)^{p^{n-1}}]}{[1-\alpha(1+p)]}.
\end{aligned}$$

Note that since $\alpha^{p^{n-1}} = 1$ either $\alpha = 1$, or $1-\alpha$ is a uniformizer in $\mathbb{Q}_p(\zeta_{p^d})$ for some $d \leq n-1$. Therefore, either $-p = [1-\alpha(1+p)]$ if $\alpha = 1$, or $[1-\alpha(1+p)]$ is a uniformizer in $\mathbb{Q}_p(\zeta_{p^d})$. Finally, it suffices to observe that

$$(1+p)^{p^{n-1}} \equiv 1 \bmod p^n;$$

hence, $\frac{1}{p^{n-1}}\sum_{b=1}^{p^{n-1}}(1+bp)\chi_2(1+bp)$ is integral.

If $\chi_1 = \omega_p^{-1}$, the same argument gives nonintegrality, and analyzing the proof gives the last assertion in the statement of the lemma regarding $B_{1,\chi}$ being a uniformizing parameter in the local field $\mathbb{Q}_p(B_{1,\chi}) = \mathbb{Q}_p(\chi)$; we omit the details. □

Corollary 6. *For p a prime, let χ be a Dirichlet character on $(\mathbb{Z}/p^n\mathbb{Z})^\times$ with $\chi(-1) = -1$. Write $(\mathbb{Z}/p^n\mathbb{Z})^\times = (\mathbb{Z}/p\mathbb{Z})^\times \times (1+p\mathbb{Z}/1+p^n\mathbb{Z})$, and the character χ as $\chi_1 \times \chi_2$ with respect to this decomposition. Then*

$$\prod_{\substack{\chi=\chi_1\times\chi_2 \\ \chi_1=\omega_p^{-1}}} L(0,\chi)$$

belongs to $\mathbb{Q}_p$, and has valuation $-n$ as an element of $\mathbb{Q}_p$.

The following proposition follows by putting the previous two lemmas together, and making an argument similar to what went into the proof of these two lemmas. We omit the details.

Proposition 7. *Primitive odd Dirichlet characters* $\chi : (\mathbb{Z}/n)^\times \to \overline{\mathbb{Z}}_p^\times$ *for which* $L(0, \chi)$ *does not belong to* $\overline{\mathbb{Z}}_p$ *are exactly those for which*

(1) $n = p^d$ *and*

(2) $\chi = \omega_p^{-1} \bmod \mathfrak{p}$.

The following consequence of the proposition suggests that prudence is to be exercised when discussing congruences of L-values for Artin representations which are congruent.

Corollary 8. *Let* p, q *be odd primes with* $p \mid (q-1)$. *For any character* χ_2 *of* $(\mathbb{Z}/q\mathbb{Z})^\times$ *of order* p, *define the character* $\chi = \omega_p^{-1} \times \chi_2$ *of* $(\mathbb{Z}/pq\mathbb{Z})^\times$. *Then although the characters* ω_p^{-1} *and* χ *have the same reduction modulo* p, $L(0, \omega_p^{-1})$ *is* p*-adically nonintegral whereas* $L(0, \chi)$ *is integral.*

Question. Let $\chi : (\mathbb{Z}/p^d m)^\times \to \overline{\mathbb{Z}}_p^\times$ with $(p, m) = 1$, $m > 1$, be a primitive Dirichlet character for which $\chi = \omega_p^{-1} \bmod \mathfrak{p}$ so that by Proposition 7, $L(0, \chi)$ is $\mathfrak{p}$-integral. Is it possible to have $L(0, \chi) = 0$ modulo $\mathfrak{p}$, the maximal ideal of $\overline{\mathbb{Z}}_p$? Our proofs in this section are "up to $\overline{\mathbb{Z}}$", so good to detect integrality, but not good for questions modulo $\mathfrak{p}$. The question is relevant to Conjecture 3 to see if the character ω_p appears in the class group H/H^+ for $E = \mathbb{Q}(\zeta_{p^d m})$; such a character is known not to appear in the class group of H/H^+ for $E = \mathbb{Q}(\zeta_{p^d})$.

5. Congruences and their failure for L-values

This paper considers integrality properties of certain Artin L-functions at 0. It may seem most natural that if two such Artin representations $\rho_1, \rho_2 : \mathrm{Gal}(\overline{\mathbb{Q}}/k) \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ have the same semisimplification mod p and do not contain the character ω_p^{-1}, then $L(0, \rho_1)$ and $L(0, \rho_2)$, which are in $\overline{\mathbb{Z}}_p$ by Conjecture 1, have the same reduction mod p. This is not true even in the simplest case of Dirichlet characters for $\mathbb{Q}$. It is possible to fix this problem for abelian characters of $\mathbb{Q}$, and more generally for any totally real number field, which is what this section strives to do; see Proposition 13. The recipe given in Proposition 13 immediately suggests itself in the nonabelian case, but we have not spelled it out.

The problem that we find dealing with abelian characters χ_1, χ_2 is that they may be congruent for some prime, but may have different conductors in which case it is not the L-values $L(0, \chi_1)$ and $L(0, \chi_2)$ which are congruent, but a modified L-value, say $L_f(0, \chi)$, which gives the right congruence; these L-values are products of $\prod_{\wp}(1 - \chi(\wp))$ with $L(0, \chi)$ where $\wp$ are all primes dividing either the conductor of χ_1 or of χ_2.

We begin with some elementary lemmas which go into congruences of L-values at 0 of Dirichlet characters, and then we consider totally real number fields.

Lemma 9. *Let V be a vector space over $\mathbb{Q}$, and $\chi = \chi_f : \mathbb{Z}/f \to V$ be any function with the property that $\sum_{a=1}^{f} \chi(a) = 0$. Let χ_{df} be the function on $\mathbb{Z}/df$ obtained from χ by composing with the natural map $\mathbb{Z}/df \to \mathbb{Z}/f$. Then*:

(1) $L(0, \chi) := \frac{1}{f} \sum_{a=1}^{f} a\chi(a) = \frac{1}{df} \sum_{a=1}^{df} a\chi_{df}(a) := L(0, \chi_{df})$.

(2) *Let $\chi : (\mathbb{Z}/f)^\times \to \mathbb{C}^\times$ be a primitive character of conductor f with $\chi \neq 1$. Then for any $f \mid f'$,*

$$\frac{1}{f'} \sum_{\substack{a=1 \\ (a,f')=1}}^{f'} a\chi(a) = \prod_{p \mid f'} (1 - \chi(p)) L(0, \chi).$$

Proof. Observe that

$$\begin{aligned}
\sum_{a=1}^{df} a\chi_{df}(a) &= \sum_{a=1}^{f} \sum_{i=0}^{d-1} (a + if)\chi_{df}(a + if) \\
&= \sum_{a=1}^{f} \sum_{i=0}^{d-1} (a + if)\chi(a) \\
&= \sum_{a=1}^{f} \left[da\chi(a) + \frac{fd(d-1)}{2} \chi(a) \right] \\
&= d \sum_{a=1}^{f} a\chi(a),
\end{aligned}$$

where in the last step we have used that $\sum_{a=1}^{f} \chi(a) = 0$. The proof of part (1) of the lemma follows.

The proof of part (2) will proceed in several steps, according to the value of f'. Observe first that if f and f' have the same prime divisors, then $(a, f) = 1$ if and only $(a, f') = 1$. Therefore,

$$\frac{1}{f'} \sum_{\substack{a=1 \\ (a,f')=1}}^{f'} a\chi(a) = \frac{1}{f'} \sum_{a=1}^{f'} a\chi_{f'}(a) = \frac{1}{f} \sum_{a=1}^{f} a\chi_f(a),$$

where the second equality is a consequence of part (1) of the lemma. In this case, i.e., when f and f' have the same prime divisors, for all $p \mid f'$, $\chi(p) = 0$. It follows that $\prod_{p \mid f'} (1 - \chi(p)) = 1$, proving this case of part (2) of the lemma.

Assume next that $f' = fp^m$, $m \geq 1$ and p a prime with $(f, p) = 1$. In this case note that (using the notation of $L(0, \chi_{fp^m})$ introduced in part (1) of the lemma),

$$L(0, \chi_{fp^m}) = \frac{1}{fp^m} \sum_{\substack{a=1 \\ (a, fp^m)=1}}^{fp^m} a\chi(a) + \frac{1}{fp^m} \sum_{i=1}^{fp^{m-1}} (pi)\chi_{fp^{m-1}}(pi)$$

$$= \frac{1}{fp^m} \sum_{\substack{a=1 \\ (a, fp^m)=1}}^{fp^m} a\chi(a) + \chi(p)L(0, \chi_{fp^{m-1}}).$$

Since by part (1) of the lemma, $L(0, \chi_{fp^m}) = L(0, \chi_{fp^{m-1}})$, the proof of part (2) follows in this case.

For general $f' = df''$, with f'' having the same prime divisors as f, and d having prime divisors which are coprime to those of f, let $d = p_1^{m_1} \cdots p_r^{m_r}$. We argue by induction on r, thus assuming the result for $d_{r-1} = p_1^{m_1} \cdots p_{r-1}^{m_{r-1}}$, adding the prime power $p_r^{m_r}$ at the end which proves (2) for $d = p_1^{m_1} \cdots p_r^{m_r}$ using part (1), and part of (2) just proved for prime powers (to be used for $p_r^{m_r}$). □

Lemma 10. *Let $\chi_1, \chi_2 : (\mathbb{Z}/f)^\times \to \overline{\mathbb{Z}}_p^\times$ be two (not necessarily primitive) odd characters. Consider χ_1, χ_2 as functions on $\mathbb{Z}/f$ by declaring their values outside of $(\mathbb{Z}/f)^\times$ to be zero. Assume that the reductions mod $\wp$, $\overline{\chi}_1, \overline{\chi}_2 : (\mathbb{Z}/f)^\times \to \overline{\mathbb{F}}_p^\times$, are the same. If $p \mid f$, assume that neither of the $\overline{\chi}_1, \overline{\chi}_2 : (\mathbb{Z}/f)^\times \to \overline{\mathbb{F}}_p^\times$ factors through $(\mathbb{Z}/p)^\times \to \overline{\mathbb{F}}_p^\times$ to give ω_p^{-1} where ω_p is the natural map from $(\mathbb{Z}/p)^\times$ to $\mathbb{F}_p^\times$. Then $L_f(0, \chi_1) := \frac{1}{f}\sum_{a=1}^f a\chi_1(a)$ and $L_f(0, \chi_2) := \frac{1}{f}\sum_{a=1}^f a\chi_2(a)$ are in $\overline{\mathbb{Z}}_p$, and have the same reduction to $\overline{\mathbb{F}}_p$.*

Proof. By the hypothesis in the lemma, there is a $b \in (\mathbb{Z}/f)^\times$ such that $[b\chi_1(b) - 1] \in \overline{\mathbb{Z}}_p^\times$, and hence also $[b\chi_2(b) - 1] \in \overline{\mathbb{Z}}_p^\times$. Fix such a $b \in (\mathbb{Z}/f)^\times$.

For $a \in \{1, \ldots, f\}$, write

$$ab = [ab] + \lambda_a f,$$

with $[ab] \in \{1, \ldots, f\}$.

From the definition of $L_f(0, \chi_1)$,

$$b\chi_1(b)L_f(0, \chi_1) = \frac{1}{f} \sum_{a=1}^{f} ab\chi_1(ab).$$

As b is invertible in $\mathbb{Z}/f$, $a \to [ab]$ is a bijection on $\{1, \ldots, f\}$; therefore, the above equation yields

$$[b\chi_1(b) - 1]L_f(0, \chi_1) = \sum_{a=1}^{f} \lambda_a \chi_1(ab). \tag{6}$$

Similarly,

$$[b\chi_2(b) - 1]L_f(0, \chi_2) = \sum_{a=1}^{a=f} \lambda_a \chi_2(ab). \tag{7}$$

Since χ_1, χ_2 are congruent, the right-hand sides of (6) and (7) are the same in $\bar{\mathbb{F}}_p$, and by the choice of b made in the beginning of the proof of the lemma, $[b\chi_1(b)-1]$ as well as $[b\chi_2(b) - 1]$ are in $\bar{\mathbb{F}}_p^\times$, and are the same; thus, it follows that $L_f(0, \chi_1)$ and $L_f(0, \chi_2)$ are in $\bar{\mathbb{Z}}_p$, and are the same in $\bar{\mathbb{F}}_p$. □

Proposition 11. *Let f_1, f_2 be integers, and f any integer divisible by both f_1, f_2. Suppose χ_1 and χ_2 are primitive odd Dirichlet characters of conductors f_1 and f_2 with values in $\bar{\mathbb{Z}}_p^\times$, respectively. If $p \mid f$, assume that neither of the $\bar{\chi}_1, \bar{\chi}_2 : (\mathbb{Z}/f)^\times \to \bar{\mathbb{F}}_p^\times$ factor through $(\mathbb{Z}/p)^\times \to \bar{\mathbb{F}}_p^\times$ to give ω_p^{-1} where ω_p is the natural map from $(\mathbb{Z}/p)^\times$ to $\mathbb{F}_p^\times$. Then $L_f(0, \chi_1) := \frac{1}{f}\sum_{a=1}^{f} a\chi_1(a)$ (where χ_1 is considered as a function on $\mathbb{Z}/f$ zero outside $(\mathbb{Z}/f)^\times$) has the value given by*

$$L_f(0, \chi_1) = \prod_{p|f} (1 - \chi_1(p)) \cdot L(0, \chi_1)$$

and similarly for $L_f(0, \chi_2)$. Both $L(0, \chi_1)$ and $L(0, \chi_2)$ are $\bar{\mathbb{Z}}_p$, and if χ_1 and χ_2 are congruent modulo the maximal ideal in $\bar{\mathbb{Z}}_p$, so is the case for $L_f(0, \chi_1)$ and $L_f(0, \chi_2)$.

Proof. $L_f(0, \chi_1) := \frac{1}{f}\sum_{a=1}^{f} a\chi_1(a)$ has the value as asserted in the proposition by part (2) of Lemma 9, and their congruence holds by Lemma 10. □

Corollary 12. *If χ_1 and χ_2 have conductors f_1 and f_2 such that the prime divisors of f_1 and f_2 are the same, then for f which is the least common multiple of f_1 and f_2, $L_f(0, \chi_1) = L(0, \chi_1)$ and $L_f(0, \chi_2) = L(0, \chi_2)$; hence, if χ_1 and χ_2 are congruent, so are $L(0, \chi_1)$ and $L(0, \chi_2)$. On the other hand, suppose $f_1 = p$, $f_2 = pq$, and $\chi_2 = \chi_1 \times \alpha : (\mathbb{Z}/p)^\times \times (\mathbb{Z}/q)^\times = (\mathbb{Z}/pq)^\times \to \bar{\mathbb{Z}}_p^\times$ with $\bar{\alpha} = 1$; then*

$$L_{pq}(0, \chi_2) = L(0, \chi_2),$$
$$L_{pq}(0, \chi_1) = (1 - \chi_1(q))L(0, \chi_1).$$

Since $L_{pq}(0, \chi_1)$ and $L_{pq}(0, \chi_2)$ are congruent mod $\wp$, we find that if $L(0, \chi_1)$ and $L(0, \chi_2)$ are not both zero mod $\wp$, they cannot be the same mod $\wp$ since $(1-\chi_1(q))$ cannot be $1 \bmod \wp$.

Remark. The hypothesis that the reduction mod $\wp$ of $\chi : (\mathbb{Z}/f)^\times \to \bar{\mathbb{Z}}_p^\times$ does not factor through $(\mathbb{Z}/p)^\times \to \bar{\mathbb{F}}_p^\times$ to give ω_p^{-1} is stronger than what is required for p-integrality of $L(0, \chi)$. For example, by Lemma 4, $L(0, \chi)$ is integral if the conductor of χ has two distinct prime factors, say $f = pq$, with $(\mathbb{Z}/f)^\times = (\mathbb{Z}/p)^\times \times (\mathbb{Z}/q)^\times$, and $\chi = \alpha \times \beta$. In Lemma 10 we would be excluding characters χ for which $\bar{\alpha} = \omega_p^{-1}$, and $\bar{\beta} = 1$.

Here is the more general version of the previous proposition.

Proposition 13. *Let k be a totally real number field with $G = \mathrm{Gal}(\overline{\mathbb{Q}}/k)$ its absolute Galois group. For $\chi : G \to \overline{\mathbb{Z}}_p^\times$, a character of finite order of G, which is also to be considered as a character of the group of ideals coprime to a nonzero ideal $\mathfrak{f}$ in k by class field theory (so $\mathfrak{f}$ is divisible by the conductor of χ, but may not be the conductor of χ), let $L(s, \chi)$ be the "true" L-function associated to the character χ, and define $L_\mathfrak{f}(s, \chi)$ by*

$$L_\mathfrak{f}(s, \chi) = \sum_{(\mathfrak{a}, \mathfrak{f})=1} \frac{\chi(\mathfrak{a})}{(N\mathfrak{a})^s},$$

where $N\mathfrak{a}$ denotes the norm of an integral ideal $\mathfrak{a}$ in k. Then:

$$L_\mathfrak{f}(s, \chi) = \prod_{\wp | \mathfrak{f}} \left(1 - \frac{\chi(\wp)}{(N\wp)^s}\right) \cdot L(s, \chi). \tag{1}$$

(2) *For any integral ideals $\mathfrak{c}$ in k coprime to $p\mathfrak{f}$, integers $k \geq 1$,*

$$\Delta_\mathfrak{c}(1-k, \chi) = (1 - \chi(\mathfrak{c})N\mathfrak{c}^k) L_\mathfrak{f}(1-k, \chi)$$

are in $\overline{\mathbb{Z}}_p$.

(3) *If χ_1 and χ_2 are two characters of G with values in $\overline{\mathbb{Z}}_p^\times$ with conductors dividing $\mathfrak{f}$, such that neither of the two reductions $\overline{\chi}_1, \overline{\chi}_2 : G \to \overline{\mathbb{F}}_p^\times$ is ω_p^{-1}, then $L_\mathfrak{f}(0, \chi_1)$ and $L_\mathfrak{f}(0, \chi_2)$ are in $\overline{\mathbb{Z}}_p$, and if χ_1 and χ_2 are congruent modulo the maximal ideal in $\overline{\mathbb{Z}}_p$, so is the case for $L_\mathfrak{f}(0, \chi_1)$ and $L_\mathfrak{f}(0, \chi_2)$.*

Proof. Deligne and Ribet [1980, Theorem 0.4] (see also [Ribet 1979, Theorem 2.1, Proposition 1.4]) prove integrality of $\Delta_\mathfrak{c}(1-k, \chi)$, as well as the congruence between $\Delta_\mathfrak{c}(1-k, \chi_1)$ and $\Delta_\mathfrak{c}(1-k, \chi_2)$.

The final congruence between $L_\mathfrak{f}(0, \chi_1)$ and $L_\mathfrak{f}(0, \chi_2)$ follows as in Lemma 10 by choosing an integral ideal $\mathfrak{c}$ coprime to $p\mathfrak{f}$ such that $(1 - \chi(\mathfrak{c})N\mathfrak{c})$ is a unit in $\overline{\mathbb{Z}}_p$ which follows just as in Lemma 10 for $\chi = \chi_1$ (hence for $\chi = \chi_2$ too), because $\overline{\chi}_1 : G \to \overline{\mathbb{F}}_p^\times$ is not ω_p^{-1} (see in the beginning of [Ribet 1979, §2] how the "norm map" under the identification of characters on ideals coprime to $\mathfrak{f}$ to characters of G becomes action of G on p-power roots of identity). □

The discussion in this section leads us to the following conjecture.

Conjecture 14. *Let k be a totally real number field and p an odd prime. Let $\bar{\rho} : \mathrm{Gal}(\overline{\mathbb{Q}}/k) \to \mathrm{GL}_n(\overline{\mathbb{F}}_p)$ be a set of semisimple modular representations of the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$ with $\bar{\rho}(c) = -1$, where c is a complex conjugation in $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$. Assume that the set of representations $\bar{\rho}$ considered are unramified outside a fixed finite set of finite places S of k, and that $\omega \otimes \bar{\rho}$ does not contain the*

trivial representation of $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$ *where* ω *is the action of* $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$ *on the* p*-th roots of unity. Then it is possible to define* $\bar{L}_S(0, \bar{\rho}) \in \overline{\mathbb{F}}_p$ *with*

$$\bar{L}_S(0, \bar{\rho}_1 + \bar{\rho}_2) = \bar{L}_S(0, \bar{\rho}_1) \cdot \bar{L}_S(0, \bar{\rho}_2),$$

for any two such representations $\bar{\rho}_1$ *and* $\bar{\rho}_2$*, and such that, if* $\bar{\rho}$ *arises as the semisimplification of reduction mod* p *of a representation* $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/k) \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ *with* $\rho(c) = -1$*, where* c *is a complex conjugation in* $\mathrm{Gal}(\overline{\mathbb{Q}}/k)$*, then*

$$L_S(0, \rho) = L(0, \rho) / \prod_{v \in S} L(0, \rho_v),$$

which belongs to $\overline{\mathbb{Z}}_p$ *by Conjecture 1, has its reduction mod* p *to be* $\bar{L}_S(0, \bar{\rho})$*.*

The conjecture above requires that if two representations $\rho_1, \rho_2 : \mathrm{Gal}(\overline{\mathbb{Q}}/k) \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ have the same semisimplification mod p, then $L_S(0, \rho_1)$ and $L_S(0, \rho_2)$ are in $\overline{\mathbb{Z}}_p$ and have the same reduction mod p. By a well known theorem of Brauer, a modular representation $\bar{\rho}$ can be lifted to a virtual representation $\sum n_i \rho_i$ in characteristic 0. However, since $L(0, \rho_i)$ may be zero mod p, for some i (for which $n_i < 0$), the theorem of Brauer does not guarantee that $\bar{L}_S(0, \bar{\rho})$ can be defined.

Acknowledgements

I thank P. Colmez for suggesting that the conjectures formulated here are not as outrageous as one might think, and for even suggesting that some of the conjectures above should have an affirmative answer as a consequence of the main conjecture of Iwasawa theory for totally real number fields proved by A. Wiles [1990] if one knew the vanishing of the μ-invariant (which is a conjecture of Iwasawa proved for abelian extensions of $\mathbb{Q}$ by Ferrero and Washington). I also thank U. K. Anandavardhanan, C. Dalawat, and C. Khare for their comments and their encouragement. I am particularly grateful to G. Asvin for his help with the last section. I also thank the referee for useful comments.

The final writing of this work was supported by a grant of the Government of the Russian Federation for the state support of scientific research carried out under the supervision of leading scientists, agreement 14.W03.31.0030, February 15, 2018.

References

[Deligne and Ribet 1980] P. Deligne and K. A. Ribet, "Values of abelian L-functions at negative integers over totally real fields", *Invent. Math.* **59**:3 (1980), 227–286. MR Zbl

[Gras 1977] G. Gras, "Étude d'invariants relatifs aux groupes des classes des corps abéliens", pp. 35–53 in *Journées Arithmétiques de Caen* (Caen, 1976), Astérisque **41–42**, Soc. Math. France, Paris, 1977. MR Zbl

[Mazur and Wiles 1984] B. Mazur and A. Wiles, "Class fields of abelian extensions of **Q**", *Invent. Math.* **76**:2 (1984), 179–330. MR Zbl

[Ribet 1976] K. A. Ribet, "A modular construction of unramified p-extensions of $Q(\mu_p)$", *Invent. Math.* **34**:3 (1976), 151–162. MR Zbl

[Ribet 1979] K. A. Ribet, "Report on p-adic L-functions over totally real fields", pp. 177–192 in *Journées Arithmétiques de Luminy* (Luminy, 1978), Astérisque **61**, Soc. Math. France, Paris, 1979. MR Zbl

[Solomon 1990] D. Solomon, "On the classgroups of imaginary abelian fields", *Ann. Inst. Fourier* (*Grenoble*) **40**:3 (1990), 467–492. MR Zbl

[Tate 1984] J. Tate, *Les conjectures de Stark sur les fonctions L d'Artin en s = 0*, Progress in Mathematics **47**, Birkhäuser, Boston, 1984. MR Zbl

[Washington 1982] L. C. Washington, *Introduction to cyclotomic fields*, Graduate Texts in Mathematics **83**, Springer, 1982. MR Zbl

[Wiles 1990] A. Wiles, "The Iwasawa conjecture for totally real fields", *Ann. of Math.* (2) **131**:3 (1990), 493–540. MR Zbl

Received March 10, 2018. Revised April 5, 2019.

DIPENDRA PRASAD
INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
MUMBAI
INDIA

and

SCHOOL OF MATHEMATICS
TATA INSTITUTE OF FUNDAMENTAL RESEARCH
MUMBAI
INDIA

and

LABORATORY OF MODERN ALGEBRA AND APPLICATIONS
SAINT PETERSBURG STATE UNIVERSITY
SAINT PETERSBURG
RUSSIA

prasad.dipendra@gmail.com

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.317

TRANSITIVE TOPOLOGICAL MARKOV CHAINS OF GIVEN ENTROPY AND PERIOD WITH OR WITHOUT MEASURE OF MAXIMAL ENTROPY

SYLVIE RUETTE

We show that, for every positive real number h and every positive integer p, there exist oriented graphs G, G' (with countably many vertices) that are strongly connected, of period p, of Gurevich entropy h, and such that G is positive recurrent (thus the topological Markov chain on G admits a measure of maximal entropy) and G' is transient (thus the topological Markov chain on G' admits no measure of maximal entropy).

1. Vere-Jones classification of graphs

In this paper, all the graphs are oriented and have a finite or countable set of vertices, and if u, v are two vertices, there is at most one arrow $u \to v$. A *path* of length n in the graph G is a sequence of vertices $(u_0, u_1, \dots, u_n)$ such that $u_i \to u_{i+1}$ in G for all $i \in [\![0, n-1]\!]$. This path is called a *loop* if $u_0 = u_n$.

Definition 1. Let G be an oriented graph, and let u, v be two vertices in G. We define the following quantities:

- $p_{uv}^G(n)$ is the number of paths $(u_0, u_1, \dots, u_n)$ such that $u_0 = u$ and $u_n = v$; $R_{uv}(G)$ is the radius of convergence of the series $\sum p_{uv}^G(n) z^n$.
- $f_{uv}^G(n)$ is the number of paths $(u_0, u_1, \dots, u_n)$ such that $u_0 = u$, $u_n = v$, and $u_i \neq v$ for all $0 < i < n$; $L_{uv}(G)$ is the radius of convergence of the series $\sum f_{uv}^G(n) z^n$.

Definition 2. Let G be an oriented graph and V its set of vertices. The graph G is *strongly connected* if, for all $u, v \in V$, there exists a path from u to v in G. The *period* of a strongly connected graph G is the greatest common divisor of $(p_{uu}^G(n))_{u \in V,\, n \geq 0}$. The graph G is *aperiodic* if its period is 1.

Proposition 3 [Vere-Jones 1962]. *Let G be an oriented graph. If G is strongly connected, $R_{uv}(G)$ does not depend on u and v; it is denoted by $R(G)$.*

MSC2010: primary 37B10; secondary 37B40.
Keywords: topological Markov chain, countable oriented graph, topological entropy.

	transient	null recurrent	positive recurrent
$\sum_{n>0} f_{uu}^G(n)R^n$	< 1	1	1
$\sum_{n>0} n f_{uu}^G(n)R^n$	$\leq +\infty$	$+\infty$	$< +\infty$
$\sum_{n\geq 0} p_{uv}^G(n)R^n$	$< +\infty$	$+\infty$	$+\infty$
$\lim_{n\to+\infty} p_{uv}^G(n)R^n$	0	0	$\lambda_{uv} > 0$
	$R = L_{uu}$	$R = L_{uu}$	$R \leq L_{uu}$

Table 1. Properties of the series associated to a transient, null recurrent or positive recurrent graph G (G is strongly connected); these properties do not depend on the vertices u, v.

If there is no confusion, $R(G)$ and $L_{uv}(G)$ will be written R and L_{uv}.

Vere-Jones [1962] gives a classification of strongly connected graphs as transient, null recurrent, or positive recurrent. These definitions are lines 1 and 2 in Table 1. The other lines of Table 1 state properties of the series $\sum p_{uv}^G(n)z^n$, which give alternative definitions (lines 3 and 4 are in [Vere-Jones 1962], and the last line is Proposition 4).

Proposition 4 [Salama 1992]. *Let G be a strongly connected oriented graph. If G is transient or null recurrent, then $R = L_{uu}$ for all vertices u. Equivalently, if there exists a vertex u such that $R < L_{uu}$, then G is positive recurrent.*

2. Topological Markov chains and Gurevich entropy

Let G be an oriented graph and V its set of vertices. We define Γ_G as the set of two-sided infinite paths in G, that is,

$$\Gamma_G := \{(v_n)_{n\in\mathbb{Z}} \mid \text{for all } n \in \mathbb{Z},\ v_n \to v_{n+1} \text{ in } G\} \subset V^{\mathbb{Z}}.$$

The map σ is the shift on Γ_G. The *topological Markov chain* on the graph G is the dynamical system (Γ_G, σ).

The set V is endowed with the discrete topology, and Γ_G is endowed with the induced topology of $V^{\mathbb{Z}}$. The space Γ_G is not compact unless G is finite.

The topological Markov chain (Γ_G, σ) is transitive if and only if the graph G is strongly connected. It is topologically mixing if and only if the graph G is strongly connected and aperiodic.

If G is a finite graph, Γ_G is compact and the topological entropy $h_{\text{top}}(\Gamma_G, \sigma)$ is well defined (see, e.g., [Denker et al. 1976] for the definition of the topological entropy). If G is a countable graph, the *Gurevich entropy* [1969] of the graph G (or of the topological Markov chain Γ_G) is given by

$$h(G) := \sup\{h_{\text{top}}(\Gamma_H, \sigma) \mid H \subset G,\ H \text{ finite}\}.$$

This entropy can also be computed in a combinatorial way, as the exponential growth of the number of paths with fixed endpoints.

Proposition 5 [Gurevich 1970]. *Let G be a strongly connected oriented graph. Then for all vertices u, v,*

$$h(G) = \lim_{n\to+\infty} \frac{1}{n} \log p_{uv}^G(n) = -\log R(G).$$

Moreover, the variational principle is still valid for topological Markov chains.

Theorem 6 [Gurevich 1969]. *Let G be an oriented graph. Then*

$$h(G) = \sup\{h_\mu(\Gamma_G) \mid \mu \ \sigma\text{-invariant probability measure}\}.$$

In this variational principle, the supremum is not necessarily reached. The next theorem gives a necessary and sufficient condition for the existence of a measure of maximal entropy (that is, a probability measure μ such that $h(G) = h_\mu(\Gamma_G)$) when the graph is strongly connected.

Theorem 7 [Gurevich 1970]. *Let G be a strongly connected oriented graph of finite positive entropy. Then the topological Markov chain on G admits a measure of maximal entropy if and only if the graph G is positive recurrent. Moreover, such a measure is unique if it exists.*

3. Construction of graphs of given entropy and given period that are either positive recurrent or transient

Lemma 8. *Let $\beta \in (1, +\infty)$. There exist a sequence of nonnegative integers $(a(n))_{n\geq 1}$ and positive constants c, M such that*

- $a(1) = 1$,
- $\sum_{n\geq 1} a(n)(1/\beta^n) = 1$,
- *for all $n \geq 2$, $c \cdot \beta^{n^2-n} \leq a(n^2) \leq c \cdot \beta^{n^2-n} + M$,*
- *for all $n \geq 1$, $0 \leq a(n) \leq M$ if n is not a square.*

These properties imply that the radius of convergence of $\sum_{n\geq 1} a(n)z^n$ is $L = 1/\beta$ and that $\sum_{n\geq 1} na(n)L^n < +\infty$.

Proof. First we look for a constant $c > 0$ such that

$$\frac{1}{\beta} + c \sum_{n\geq 2} \beta^{n^2-n} \frac{1}{\beta^{n^2}} = 1. \tag{1}$$

We have

$$\sum_{n\geq 2} \beta^{n^2-n} \frac{1}{\beta^{n^2}} = \sum_{n\geq 2} \beta^{-n} = \frac{1}{\beta(\beta-1)}.$$

Thus,

$$(1) \quad \Longleftrightarrow \quad \frac{1}{\beta}+\frac{c}{\beta(\beta-1)}=1 \quad \Longleftrightarrow \quad c=(\beta-1)^2.$$

Since $\beta>1$, the constant $c:=(\beta-1)^2$ is positive. We define the sequence $(b(n))_{n\geq 1}$ by

- $b(1):=1$,
- $b(n^2):=\lfloor c\beta^{n^2-n}\rfloor$ for all $n\geq 2$,
- $b(n):=0$ for all $n\geq 2$ such that n is not a square.

Then

$$\sum_{n\geq 1} b(n)\frac{1}{\beta^n}\leq \frac{1}{\beta}+c\sum_{n\geq 2}\beta^{n^2-n}\frac{1}{\beta^{n^2}}=1.$$

We set $\delta:=1-\sum_{n\geq 1} b(n)(1/\beta^n)\in[0,1)$ and $k:=\lfloor\beta^2\delta\rfloor$. Then $k\leq\beta^2\delta<k+1<k+\beta$, which implies that $0\leq\delta-k/\beta^2<1/\beta$. We write the β-expansion of $\delta-k/\beta^2$ (see, e.g., [Dajani and Kraaikamp 2002, p. 51] for the definition): there exist integers $d(n)\in\{0,\ldots,\lfloor\beta\rfloor\}$ such that $\delta-k/\beta^2=\sum_{n\geq 1} d(n)(1/\beta^n)$. Moreover, $d(1)=0$ because $\delta-k/\beta^2<1/\beta$. Thus, we can write

$$\delta=\sum_{n\geq 2} d'(n)\frac{1}{\beta^n}$$

where $d'(2):=d(2)+k$ and $d'(n):=d(n)$ for all $n\geq 3$.

We set $a(1):=b(1)$ and $a(n):=b(n)+d'(n)$ for all $n\geq 2$. Let $M:=\beta+k$. We then have

- $a(1)=1$,
- $\sum_{n\geq 1} a(n)(1/\beta^n)=1$,
- for all $n\geq 2$, $c\cdot\beta^{n^2-n}\leq a(n^2)\leq c\cdot\beta^{n^2-n}+\beta\leq c\cdot\beta^{n^2-n}+M$,
- $0\leq a(2)\leq\beta+k=M$,
- for all $n\geq 3$, $0\leq a(n)\leq\beta\leq M$ if n is not a square.

The radius of convergence L of $\sum_{n\geq 1} a(n)z^n$ satisfies

$$-\log L=\limsup_{n\to+\infty}\frac{1}{n}\log a(n)=\lim_{n\to+\infty}\frac{1}{n^2}\log a(n^2)=\log\beta$$

because $a(n^2)\sim c\beta^{n^2-n}$.

Thus, $L=1/\beta$. Moreover,

$$\sum_{n\geq 1} na(n)\frac{1}{\beta^n}\leq M\sum_{n\geq 1} n\frac{1}{\beta^n}+c\sum_{n\geq 1} n^2\beta^{n^2-n}\frac{1}{\beta^{n^2}}=M\sum_{n\geq 1}\frac{n}{\beta^n}+c\sum_{n\geq 1}\frac{n^2}{\beta^n}<+\infty. \ \square$$

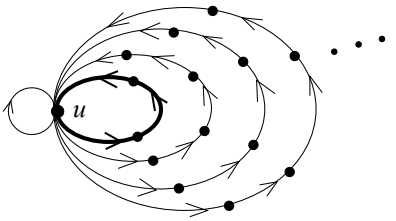


Figure 1. The graphs $G(\beta)$ and $G'(\beta)$; the bold loop belongs to $G(\beta)$ and not to $G'(\beta)$, otherwise the two graphs coincide.

Lemma 9 [Ruette 2003, Lemma 2.4]. *Let G be a strongly connected oriented graph and u a vertex.*

(i) $R < L_{uu}$ *if and only if* $\sum_{n\geq 1} f_{uu}^G(n)L_{uu}^n > 1$.

(ii) *If G is recurrent, then R is the unique positive number x such that*

$$\sum_{n\geq 1} f_{uu}^G(n)x^n = 1.$$

Proof. For (i) and (ii), use Table 1 and the fact that $F(x) = \sum_{n\geq 1} f_{uu}^G(n)x^n$ is increasing for $x \in [0, +\infty)$. □

Proposition 10. *Let $\beta \in (1, +\infty)$. There exist aperiodic strongly connected graphs $G'(\beta) \subset G(\beta)$ such that $h(G(\beta)) = h(G'(\beta)) = \log\beta$, $G(\beta)$ is positive recurrent, and $G'(\beta)$ is transient.*

Remark. Salama [1988, Theorem 3.9] proved the part of this proposition concerning positive recurrent graphs.

Proof. This is a variant of the proof of [Ruette 2003, Example 2.9].

Let u be a vertex, and let $(a(n))_{n\geq 1}$ be the sequence given by Lemma 8 for β. The graph $G(\beta)$ is composed of $a(n)$ loops of length n based at the vertex u for all $n \geq 1$ (see Figure 1). More precisely, define the set of vertices of $G(\beta)$ as

$$V := \{u\} \cup \bigcup_{n=1}^{+\infty} \{v_k^{n,i} \mid i \in [\![1, a(n)]\!],\ k \in [\![1, n-1]\!]\},$$

where the vertices $v_k^{n,i}$ above are distinct. Let $v_0^{n,i} = v_n^{n,i} = u$ for all $i \in [\![1, a(n)]\!]$. There is an arrow $v_k^{n,i} \to v_{k+1}^{n,i}$ for all $k \in [\![0, n-1]\!]$, $i \in [\![1, a(n)]\!]$, and $n \geq 2$; there is an arrow $u \to u$; and there is no other arrow in $G(\beta)$. The graph $G(\beta)$ is strongly connected, and $f_{uu}^{G(\beta)}(n) = a(n)$ for all $n \geq 1$.

By Lemma 8, the sequence $(a(n))_{n\geq 1}$ is defined such that $L = 1/\beta$ and

$$\sum_{n\geq 1} a(n)L^n = 1, \tag{2}$$

where $L = L_{uu}(G(\beta))$ is the radius of convergence of the series $\sum a(n)z^n$. If $G(\beta)$ is transient, then $R(G(\beta)) = L_{uu}(G(\beta))$ by Proposition 4. But (2) contradicts the definition of transient (see the first line of Table 1). Thus, $G(\beta)$ is recurrent, and $R(G(\beta)) = L$ by (2) and Lemma 9(ii). Moreover,

$$\sum_{n\geq 1} na(n)L^n < +\infty$$

by Lemma 8, and thus the graph $G(\beta)$ is positive recurrent (see Table 1). By Proposition 5, $h(G(\beta)) = -\log R(G(\beta)) = \log \beta$.

The graph $G'(\beta)$ is obtained from $G(\beta)$ by deleting a loop starting at u of length n_0 for some $n_0 \geq 2$ such that $a(n_0) \geq 1$ (such an integer n_0 exists because $L < +\infty$). Obviously one has $L_{uu}(G'(\beta)) = L$ and

$$\sum_{n\geq 1} f_{uu}^{G'(\beta)}(n)L^n = 1 - L^{n_0} < 1.$$

Since $R(G'(\beta)) \leq L_{uu}(G'(\beta))$, this implies that $G'(\beta)$ is transient. Moreover, $R(G'(\beta)) = L_{uu}(G'(\beta))$ by Proposition 4, so $R(G'(\beta)) = R(G(\beta))$, and hence $h(G'(\beta)) = h(G(\beta))$ by Proposition 5. Finally, both $G(\beta)$ and $G'(\beta)$ are of period 1 because of the arrow $u \to u$. □

Corollary 11. *Let p be a positive integer and $h \in (0, +\infty)$. There exist strongly connected graphs G, G' of period p such that $h(G) = h(G') = h$, G is positive recurrent, and G' is transient.*

Proof. For G, we start from the graph $G(\beta)$ given by Proposition 10 with $\beta = e^{hp}$. Let V denote the set of vertices of $G(\beta)$. The set of vertices of G is $V \times [\![1, p]\!]$, and the arrows in G are

- $(v, i) \to (v, i + 1)$ if $v \in V$ and $i \in [\![1, p - 1]\!]$,
- $(v, p) \to (w, 1)$ if $v, w \in V$ and $v \to w$ is an arrow in $G(\beta)$.

According to the properties of $G(\beta)$, G is strongly connected, of period p, and positive recurrent. Moreover, $h(G) = (1/p)h(G(\beta)) = (1/p)\log \beta = h$.

For G', we do the same starting with $G'(\beta)$. □

According to Theorem 7, the graphs of Corollary 11 satisfy that the topological Markov chain on G admits a measure of maximal entropy whereas the topological Markov chain on G' admits no measure of maximal entropy; both are transitive, of Gurevich entropy h, and supported by a graph of period p.

References

[Dajani and Kraaikamp 2002] K. Dajani and C. Kraaikamp, *Ergodic theory of numbers*, Carus Mathematical Monographs **29**, Mathematical Association of America, Washington, DC, 2002. MR Zbl

[Denker et al. 1976] M. Denker, C. Grillenberger, and K. Sigmund, *Ergodic theory on compact spaces*, Lecture Notes in Mathematics **527**, Springer, 1976. MR Zbl

[Gurevich 1969] B. M. Gurevich, "Topological entropy of a countable Markov chain", *Dokl. Akad. Nauk SSSR* **187**:4 (1969), 715–718. In Russian; translated in *Soviet Math. Dokl.* **10** (1969), 911–915. MR Zbl

[Gurevich 1970] B. M. Gurevich, "Shift entropy and Markov measures in the space of paths of a countable graph", *Dokl. Akad. Nauk SSSR* **192**:5 (1970), 963–965. In Russian; translated in *Soviet Math. Dokl.* **11** (1970), 744–747. MR Zbl

[Ruette 2003] S. Ruette, "On the Vere-Jones classification and existence of maximal measures for countable topological Markov chains", *Pacific J. Math.* **209**:2 (2003), 366–380. MR Zbl

[Salama 1988] I. A. Salama, "Topological entropy and recurrence of countable chains", *Pacific J. Math.* **134**:2 (1988), 325–341. Correction in **140**:2 (1989), 397. MR Zbl

[Salama 1992] I. A. Salama, "On the recurrence of countable topological Markov chains", pp. 349–360 in *Symbolic dynamics and its applications* (New Haven, CT, 1991), edited by P. Walters, Contemp. Math. **135**, Amer. Math. Soc., Providence, RI, 1992. MR Zbl

[Vere-Jones 1962] D. Vere-Jones, "Geometric ergodicity in denumerable Markov chains", *Quart. J. Math. Oxford Ser.* (2) **13** (1962), 7–28. MR Zbl

Received June 26, 2018.

SYLVIE RUETTE
LABORATOIRE DE MATHÉMATIQUES D'ORSAY, UMR 8628
UNIVERSITÉ PARIS-SACLAY
ORSAY
FRANCE

sylvie.ruette@universite-paris-saclay.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.325

RESTRICTED SUM FORMULA FOR FINITE AND SYMMETRIC MULTIPLE ZETA VALUES

HIDEKI MURAHARA AND SHINGO SAITO

The sum formula for finite and symmetric multiple zeta values, established by Wakabayashi and the authors, implies that if the weight and depth are fixed and the specified component is required to be more than one, then the values sum up to a rational multiple of the analogue of the Riemann zeta value. We prove that the result remains true if we further demand that the component should be more than two or that another component should also be more than one.

1. Introduction

The *multiple zeta values* and *multiple zeta-star values* are the real numbers defined by

$$\zeta(k_1,\dots,k_r)=\sum_{m_1>\cdots>m_r\geq 1}\frac{1}{m_1^{k_1}\cdots m_r^{k_r}},$$
$$\zeta^\star(k_1,\dots,k_r)=\sum_{m_1\geq\cdots\geq m_r\geq 1}\frac{1}{m_1^{k_1}\cdots m_r^{k_r}}$$

for $k_1,\dots,k_r\in\mathbb{Z}_{\geq 1}$ with $k_1\geq 2$. They are generalisations of the values of the Riemann zeta function at positive integers, and they are known to have interesting algebraic structures due to the many relations among them, the simplest being $\zeta(2,1)=\zeta(3)$. See, for example, the book by Zhao [2016] for further details on multiple zeta(-star) values.

The variants of multiple zeta values that we shall be looking at in this paper are *finite multiple zeta values* $\zeta_{\mathcal{A}}(k_1,\dots,k_r)$ and *symmetric multiple zeta values* $\zeta_{\mathcal{S}}(k_1,\dots,k_r)$ (the latter also known as symmetrised multiple zeta values and finite real multiple zeta values), both introduced by Kaneko and Zagier [$\geq$ 2019] (see [Zhao 2016] for details). Set $\mathcal{A}=\prod_p \mathbb{F}_p/\bigoplus_p \mathbb{F}_p$, where p runs over all primes.

MSC2010: primary 11M32; secondary 05A19.
Keywords: finite multiple zeta values, symmetric multiple zeta values, symmetrised multiple zeta values, finite real multiple zeta values, sum formula, restricted sum formula.

For $k_1, \dots, k_r \in \mathbb{Z}_{\geq 1}$, we define

$$\zeta_{\mathcal{A}}(k_1, \dots, k_r) = \left(\sum_{p > m_1 > \cdots > m_r \geq 1} \frac{1}{m_1^{k_1} \cdots m_r^{k_r}} \bmod p \right)_p \in \mathcal{A},$$

$$\zeta^{\star}_{\mathcal{A}}(k_1, \dots, k_r) = \left(\sum_{p > m_1 \geq \cdots \geq m_r \geq 1} \frac{1}{m_1^{k_1} \cdots m_r^{k_r}} \bmod p \right)_p \in \mathcal{A}.$$

Let $\mathcal{Z}$ denote the $\mathbb{Q}$-linear subspace of $\mathbb{R}$ spanned by the multiple zeta values. For $k_1, \dots, k_r \in \mathbb{Z}_{\geq 1}$, we define

$$\zeta_{\mathcal{S}}(k_1, \dots, k_r) = \sum_{j=0}^{r} (-1)^{k_1 + \cdots + k_j} \zeta(k_j, \dots, k_1) \zeta(k_{j+1}, \dots, k_r) \bmod \zeta(2) \in \mathcal{Z}/\zeta(2)\mathcal{Z},$$

$$\zeta^{\star}_{\mathcal{S}}(k_1, \dots, k_r) = \sum_{j=0}^{r} (-1)^{k_1 + \cdots + k_j} \zeta^{\star}(k_j, \dots, k_1) \zeta^{\star}(k_{j+1}, \dots, k_r) \bmod \zeta(2) \in \mathcal{Z}/\zeta(2)\mathcal{Z},$$

where we set $\zeta(\varnothing) = \zeta^{\star}(\varnothing) = 1$. The multiple zeta(-star) values that appear in the definition of the symmetric multiple zeta(-star) values are the regularised values if the first component is 1; although there are two ways of regularisation, called the harmonic regularisation and the shuffle regularisation, it is known that the symmetric multiple zeta values remain unchanged as elements of $\mathcal{Z}/\zeta(2)\mathcal{Z}$ no matter which regularisation we use (see [Kaneko and Zagier $\geq$ 2019]).

Kaneko and Zagier [$\geq$ 2019] made a striking conjecture that the finite multiple zeta values and the symmetric multiple zeta values are isomorphic; more precisely, if we let $\mathcal{Z}_{\mathcal{A}}$ denote the $\mathbb{Q}$-linear subspace of $\mathcal{A}$ spanned by the finite multiple zeta values, then $\mathcal{Z}_{\mathcal{A}}$ and $\mathcal{Z}/\zeta(2)\mathcal{Z}$ are isomorphic as $\mathbb{Q}$-algebras via the correspondence $\zeta_{\mathcal{A}}(k_1, \dots, k_r) \leftrightarrow \zeta_{\mathcal{S}}(k_1, \dots, k_r)$. It means that $\zeta_{\mathcal{A}}(k_1, \dots, k_r)$ and $\zeta_{\mathcal{S}}(k_1, \dots, k_r)$ satisfy the same relations, and a notable example of such relations is the sum formula (Theorem 1.1). In what follows, we use the letter $\mathcal{F}$ when it can be replaced with either $\mathcal{A}$ or $\mathcal{S}$; for example, by $\zeta_{\mathcal{F}}(1) = 0$ we mean that both $\zeta_{\mathcal{A}}(1) = 0$ and $\zeta_{\mathcal{S}}(1) = 0$ are true. We write

$$\mathfrak{Z}_{\mathcal{F}}(k) = \begin{cases} (B_{p-k}/k \bmod p)_p & \text{if } \mathcal{F} = \mathcal{A}, \\ \zeta(k) \bmod \zeta(2) & \text{if } \mathcal{F} = \mathcal{S} \end{cases}$$

for $k \in \mathbb{Z}_{\geq 2}$, where B_n denotes the n-th Bernoulli number. Note that it can be verified rather easily that

$$\zeta_{\mathcal{F}}(k-1, 1) = \mathfrak{Z}_{\mathcal{F}}(k) \quad \text{for } k \in \mathbb{Z}_{\geq 2},$$

so that $(B_{p-k}/k \bmod p)_p$ corresponds to $\zeta(k) \bmod \zeta(2)$ via the above-mentioned isomorphism $\mathcal{Z}_{\mathcal{A}} \cong \mathcal{Z}/\zeta(2)\mathcal{Z}$.

Theorem 1.1 [Saito and Wakabayashi 2015; Murahara 2016]. *For $k, r, i \in \mathbb{Z}$ with $1 \le i \le r \le k-1$, we have*

$$\sum_{\substack{k_1+\cdots+k_r=k\\ k_i\ge 2}} \zeta_{\mathcal{F}}(k_1,\dots,k_r) = (-1)^r \sum_{\substack{k_1+\cdots+k_r=k\\ k_i\ge 2}} \zeta^{\star}_{\mathcal{F}}(k_1,\dots,k_r)$$
$$= (-1)^{i-1}\left(\binom{k-1}{i-1} + (-1)^r\binom{k-1}{r-i}\right)\mathfrak{Z}_{\mathcal{F}}(k).$$

The theorem implies that the sums belong to $\mathbb{Q}\mathfrak{Z}_{\mathcal{F}}(k)$. Our main theorem states that similar sums also belong to $\mathbb{Q}\mathfrak{Z}_{\mathcal{F}}(k)$ if k is odd:

Theorem 1.2 (main theorem). *Let k be an odd integer with $k \ge 3$, and let r be an integer with $1 \le r \le k-2$.*

(1) *For $i \in \mathbb{Z}$ with $1 \le i \le r$, we have*

$$\sum_{\substack{k_1+\cdots+k_r=k\\ k_i\ge 3}} \zeta_{\mathcal{F}}(k_1,\dots,k_r) = (-1)^r \sum_{\substack{k_1+\cdots+k_r=k\\ k_i\ge 3}} \zeta^{\star}_{\mathcal{F}}(k_1,\dots,k_r) \in \mathbb{Q}\mathfrak{Z}_{\mathcal{F}}(k).$$

(2) *For distinct $i, j \in \mathbb{Z}$ with $1 \le i, j \le r$, we have*

$$\sum_{\substack{k_1+\cdots+k_r=k\\ k_i,k_j\ge 2}} \zeta_{\mathcal{F}}(k_1,\dots,k_r) = (-1)^r \sum_{\substack{k_1+\cdots+k_r=k\\ k_i,k_j\ge 2}} \zeta^{\star}_{\mathcal{F}}(k_1,\dots,k_r) \in \mathbb{Q}\mathfrak{Z}_{\mathcal{F}}(k).$$

The rational coefficients can be written explicitly, though in a rather complicated manner, in terms of binomial coefficients (see Theorem 3.1 for the precise statement).

Remark 1.3. If k is even, then $\mathfrak{Z}_{\mathcal{F}}(k) = 0$ and numerical experiments suggest that the sums are not always equal to 0.

2. Preliminary lemmas

This section will give a few preliminary lemmas that will be used to prove our main theorem in the next section.

An *index* is a (possibly empty) sequence of positive integers. For an index $\boldsymbol{k} = (k_1, \dots, k_r)$, the number r is called its *depth* and $k_1 + \cdots + k_r$ its *weight*.

Proposition 2.1. *If $(k_1, \dots, k_r)$ is a nonempty index, then*

$$\sum_{\sigma\in\mathfrak{S}_r} \zeta_{\mathcal{F}}(k_{\sigma(1)},\dots,k_{\sigma(r)}) = \sum_{\sigma\in\mathfrak{S}_r} \zeta^{\star}_{\mathcal{F}}(k_{\sigma(1)},\dots,k_{\sigma(r)}) = 0,$$

where $\mathfrak{S}_r$ denotes the symmetric group of order r.

Proof. Roughly speaking, the sums are zero because they can be written as polynomials of the values $\zeta_{\mathcal{F}}(k)$, which are all zero. For details, see [Hoffman 2015, Theorem 2.3; Saito 2017, Proposition 2.7], for example. □

We write $\{k\}^r$ for the r times repetition of k.

Corollary 2.2. *For $k, r \in \mathbb{Z}_{\geq 1}$, we have*

$$\zeta_{\mathcal{F}}(\{k\}^r) = \zeta^{\star}_{\mathcal{F}}(\{k\}^r) = 0.$$

Proof. Apply Proposition 2.1 to $(k_1, \dots, k_r) = (\{k\}^r)$. □

Definition 2.3. For each index $\boldsymbol{k}$, write its components as sums of ones, and define its *Hoffman dual* $\boldsymbol{k}^{\vee}$ as the index obtained by swapping plus signs and commas.

Example 2.4. If $\boldsymbol{k} = (2, 1, 3) = (1+1, 1, 1+1+1)$, then $\boldsymbol{k}^{\vee} = (1, 1+1+1, 1, 1) = (1, 3, 1, 1)$.

The following theorem, known as *duality*, was proved by Hoffman [2015] for the $\mathcal{F} = \mathcal{A}$ case and by Jarossay [2014] for the $\mathcal{F} = \mathcal{S}$ case:

Theorem 2.5 [Hoffman 2015; Jarossay 2014]. *If $\boldsymbol{k}$ is a nonempty index, then*

$$\zeta^{\star}_{\mathcal{F}}(\boldsymbol{k}^{\vee}) = -\zeta^{\star}_{\mathcal{F}}(\boldsymbol{k}).$$

For indices $\boldsymbol{k}$ and $\boldsymbol{l}$ of the same weight, we write $\boldsymbol{k} \preceq \boldsymbol{l}$ to mean that, writing their components as sums of ones, we can obtain $\boldsymbol{l}$ from $\boldsymbol{k}$ by replacing some (possibly none) of the plus signs with commas. For example, $(2, 1, 3) = (1+1, 1, 1+1+1) \preceq (1, 1, 1, 1+1, 1) = (1, 1, 1, 2, 1)$.

Corollary 2.6. *If $\boldsymbol{k}$ is a nonempty index of depth r, then*

$$(-1)^r \zeta_{\mathcal{F}}(\boldsymbol{k}) = \sum_{\boldsymbol{l} \succeq \boldsymbol{k}} \zeta_{\mathcal{F}}(\boldsymbol{l}).$$

Proof. An easy combinatorial argument shows that this corollary is equivalent to Theorem 2.5; see [Saito 2017, Corollary 2.15] for details. □

We adopt the standard convention for binomial coefficients that $\binom{a}{b} = 0$ if $a \in \mathbb{Z}_{\geq 0}$ and $b \in \mathbb{Z} \setminus \{0, \dots, a\}$. For notational simplicity, we write

$$\begin{bmatrix} a \\ b \end{bmatrix} = (-1)^b \binom{a}{b}$$

for $a \in \mathbb{Z}_{\geq 0}$ and $b \in \mathbb{Z}$ (not to be confused with the Stirling numbers of the first kind). Then Theorem 1.1 can be rewritten as follows:

Theorem 2.7 (another form of Theorem 1.1). *For $k, r, i \in \mathbb{Z}$ with $1 \leq i \leq r \leq k-1$, we have*

$$\begin{aligned} \sum_{\substack{k_1+\cdots+k_r=k \\ k_i \geq 2}} \zeta_{\mathcal{F}}(k_1, \dots, k_r) &= (-1)^r \sum_{\substack{k_1+\cdots+k_r=k \\ k_i \geq 2}} \zeta^{\star}_{\mathcal{F}}(k_1, \dots, k_r) \\ &= \left(\begin{bmatrix} k-1 \\ i-1 \end{bmatrix} - \begin{bmatrix} k-1 \\ r-i \end{bmatrix} \right) \mathfrak{Z}_{\mathcal{F}}(k). \end{aligned}$$

Lemma 2.8. *For $a, b \in \mathbb{Z}_{\geq 0}$ with $a+b$ odd, we have*

$$\zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^b) = -\begin{bmatrix} a+b+2 \\ a+1 \end{bmatrix} \mathfrak{Z}_{\mathcal{F}}(a+b+2) = \begin{bmatrix} a+b+2 \\ b+1 \end{bmatrix} \mathfrak{Z}_{\mathcal{F}}(a+b+2).$$

Proof. Applying Theorem 2.7 to $k = a+b+2$, $r = a+b+1$, and $i = a+1$ gives

$$\zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^b) = \left(\begin{bmatrix} a+b+1 \\ a \end{bmatrix} - \begin{bmatrix} a+b+1 \\ b \end{bmatrix} \right) \mathfrak{Z}_{\mathcal{F}}(a+b+2),$$

and we have

$$\begin{aligned} \begin{bmatrix} a+b+1 \\ a \end{bmatrix} - \begin{bmatrix} a+b+1 \\ b \end{bmatrix} &= (-1)^a \binom{a+b+1}{a} - (-1)^b \binom{a+b+1}{b} \\ &= -(-1)^{a+1} \left(\binom{a+b+1}{a} + \binom{a+b+1}{a+1} \right) \\ &= -(-1)^{a+1} \binom{a+b+2}{a+1} \\ &= -\begin{bmatrix} a+b+2 \\ a+1 \end{bmatrix}. \end{aligned}$$

By a similar reasoning, we also have

$$\begin{bmatrix} a+b+1 \\ a \end{bmatrix} - \begin{bmatrix} a+b+1 \\ b \end{bmatrix} = \begin{bmatrix} a+b+2 \\ b+1 \end{bmatrix}. \qquad \square$$

Lemma 2.9. *For $a, b \in \mathbb{Z}_{\geq 0}$ and $c \in \mathbb{Z}_{\geq -1}$ with $a+b+c$ odd, we have*

$$\zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^c, 2, \{1\}^b) = \tfrac{1}{2} \left(\begin{bmatrix} a+b+c+4 \\ a+1 \end{bmatrix} - \begin{bmatrix} a+b+c+4 \\ b+1 \end{bmatrix} \right) \mathfrak{Z}_{\mathcal{F}}(a+b+c+4),$$

where we understand that $\zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^{-1}, 2, \{1\}^b) = \zeta_{\mathcal{F}}(\{1\}^a, 3, \{1\}^b)$.

Proof. Keeping Corollary 2.2 in mind, we apply Corollary 2.6 to

$$\boldsymbol{k} = (\{1\}^a, 2, \{1\}^c, 2, \{1\}^b)$$

to get

$$\begin{aligned} &-\zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^c, 2, \{1\}^b) \\ &\quad = \zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^c, 2, \{1\}^b) + \zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^{b+c+2}) + \zeta_{\mathcal{F}}(\{1\}^{a+c+2}, 2, \{1\}^b), \end{aligned}$$

no matter whether $c = -1$ or $c \geq 0$. This, together with Lemma 2.8, gives

$$\begin{aligned} &\zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^c, 2, \{1\}^b) \\ &\qquad = -\tfrac{1}{2} \left(\zeta_{\mathcal{F}}(\{1\}^a, 2, \{1\}^{b+c+2}) + \zeta_{\mathcal{F}}(\{1\}^{a+c+2}, 2, \{1\}^b) \right) \\ &\qquad = \frac{1}{2} \left(\begin{bmatrix} a+b+c+4 \\ a+1 \end{bmatrix} - \begin{bmatrix} a+b+c+4 \\ b+1 \end{bmatrix} \right) \mathfrak{Z}_{\mathcal{F}}(a+b+c+4). \qquad \square \end{aligned}$$

3. Proof of the main theorem

Throughout this section, let k be an odd integer with $k \geq 3$, and let r, i, j be integers with $1 \leq i \leq j \leq r \leq k-2$. Set

$$I_{k,r,i,j} = \begin{cases} \{(k_1, \dots, k_r) \in \mathbb{Z}^r_{\geq 1} \mid k_i \geq 3\} & \text{if } i = j, \\ \{(k_1, \dots, k_r) \in \mathbb{Z}^r_{\geq 1} \mid k_i, k_j \geq 2\} & \text{if } i < j, \end{cases}$$

and write

$$S_{k,r,i,j} = \sum_{\boldsymbol{k} \in I_{k,r,i,j}} \zeta_{\mathcal{F}}(\boldsymbol{k}), \quad S^{\star}_{k,r,i,j} = \sum_{\boldsymbol{k} \in I_{k,r,i,j}} \zeta^{\star}_{\mathcal{F}}(\boldsymbol{k}).$$

For notational simplicity, we put $i' = j - i + 1$, $i'' = r - j + 1$, and $k' = k - r - 2$, so that $i + i' + i'' + k' = k$.

The aim of this section is to prove the following theorem, from which Theorem 1.2 easily follows:

Theorem 3.1. *We have*

$$S_{k,r,i,j} = (-1)^r S^{\star}_{k,r,i,j} = \tfrac{1}{2} N_{k,r,i,j} \mathfrak{Z}_{\mathcal{F}}(k),$$

where $N_{k,r,i,j}$ is an integer given by

$$\begin{aligned} N_{k,r,i,j} = {} & (k'+i+1)\left(\begin{bmatrix} k-1 \\ k'+i \end{bmatrix} - \begin{bmatrix} k-1 \\ i-1 \end{bmatrix}\right) - (k'+i''+1)\left(\begin{bmatrix} k-1 \\ k'+i'' \end{bmatrix} - \begin{bmatrix} k-1 \\ i''-1 \end{bmatrix}\right) \\ & + k\left(\begin{bmatrix} k-2 \\ k'+i-1 \end{bmatrix} - \begin{bmatrix} k-2 \\ i-2 \end{bmatrix} - \begin{bmatrix} k-2 \\ k'+i''-1 \end{bmatrix} + \begin{bmatrix} k-2 \\ i''-2 \end{bmatrix}\right). \end{aligned}$$

Proof that $S_{k,r,i,j} = (-1)^r S^{\star}_{k,r,i,j}$. In this subsection, we shall prove that $S_{k,r,i,j} = (-1)^r S^{\star}_{k,r,i,j}$ (Lemma 3.4).

Proposition 3.2. *If $(k_1, \dots, k_r)$ is an index, then*

$$\begin{aligned} \zeta_{\mathcal{F}}(k_r, \dots, k_1) &= (-1)^{k_1 + \cdots + k_r} \zeta_{\mathcal{F}}(k_1, \dots, k_r), \\ \zeta^{\star}_{\mathcal{F}}(k_r, \dots, k_1) &= (-1)^{k_1 + \cdots + k_r} \zeta^{\star}_{\mathcal{F}}(k_1, \dots, k_r). \end{aligned}$$

Proof. Easy from the definitions; see [Saito 2017, Proposition 2.6] for details. □

Proposition 3.3. *If $\boldsymbol{k} = (k_1, \dots, k_r)$ is a nonempty index, then*

$$\sum_{s=0}^{r} (-1)^s \zeta^{\star}_{\mathcal{F}}(k_s, \dots, k_1) \zeta_{\mathcal{F}}(k_{s+1}, \dots, k_r) = 0,$$

where we set $\zeta_{\mathcal{F}}(\varnothing) = \zeta^{\star}_{\mathcal{F}}(\varnothing) = 1$.

Proof. Well known; see [Saito 2017, Proposition 2.9] for the detailed proof. □

Lemma 3.4. *We have*

$$S_{k,r,i,j} = (-1)^r S^{\star}_{k,r,i,j}.$$

Proof. Adding the equation in Proposition 3.3 for all $(k_1, \dots, k_r) \in I_{k,r,i,j}$ gives

$$\sum_{s=0}^{r}(-1)^s \sum_{(k_1,\dots,k_r)\in I_{k,r,i,j}} \zeta^\star_{\mathcal{F}}(k_s, \dots, k_1)\zeta_{\mathcal{F}}(k_{s+1}, \dots, k_r) = 0,$$

whose left-hand side we shall write as $\sum_{s=0}^{r}(-1)^s A_s$ for simplicity. Observe that $A_0 = S_{k,r,i,j}$ and that

$$\begin{aligned} A_r &= \sum_{(k_1,\dots,k_r)\in I_{k,r,i,j}} \zeta^\star_{\mathcal{F}}(k_r, \dots, k_1) \\ &= \sum_{(k_1,\dots,k_r)\in I_{k,r,i,j}} (-1)^{k_1+\cdots+k_r}\zeta^\star_{\mathcal{F}}(k_1, \dots, k_r) \\ &= -S^\star_{k,r,i,j} \end{aligned}$$

by Proposition 3.2 because k is odd. For $s = j, \dots, r-1$, we have

$$\begin{aligned} A_s &= \sum_{l=0}^{k}\Bigg(\sum_{(k_1,\dots,k_s)\in I_{l,s,i,j}} \zeta^\star_{\mathcal{F}}(k_s, \dots, k_1)\Bigg)\Bigg(\sum_{k_{s+1}+\cdots+k_r=k-l} \zeta_{\mathcal{F}}(k_{s+1}, \dots, k_r)\Bigg) \\ &= 0 \end{aligned}$$

because of Proposition 2.1; we similarly have $A_s = 0$ for $s = 1, \dots, i-1$. If $i < j$ and $i \le s \le j-1$, then we have

$$\begin{aligned} A_s &= \sum_{l=0}^{k}\Bigg(\sum_{\substack{k_1+\cdots+k_s=l\\ k_i\ge 2}} \zeta^\star_{\mathcal{F}}(k_s, \dots, k_1)\Bigg)\Bigg(\sum_{\substack{k_{s+1}+\cdots+k_r=k-l\\ k_j\ge 2}} \zeta_{\mathcal{F}}(k_{s+1}, \dots, k_r)\Bigg) \\ &= \sum_{l=0}^{k}\Bigg((-1)^l\sum_{\substack{k_1+\cdots+k_s=l\\ k_i\ge 2}} \zeta^\star_{\mathcal{F}}(k_1, \dots, k_s)\Bigg)\Bigg(\sum_{\substack{k_{s+1}+\cdots+k_r=k-l\\ k_j\ge 2}} \zeta_{\mathcal{F}}(k_{s+1}, \dots, k_r)\Bigg) \\ &= \sum_{l=0}^{k}(-1)^{l+s}\left(\begin{bmatrix} l-1 \\ i-1 \end{bmatrix} - \begin{bmatrix} l-1 \\ s-i \end{bmatrix}\right)\mathfrak{Z}_{\mathcal{F}}(l)\left(\begin{bmatrix} k-l-1 \\ j-s-1 \end{bmatrix} - \begin{bmatrix} k-l-1 \\ r-j \end{bmatrix}\right)\mathfrak{Z}_{\mathcal{F}}(k-l) \end{aligned}$$

by Proposition 3.2 and Theorem 2.7; since k is odd, either l or $k-l$ must be even and so $\mathfrak{Z}_{\mathcal{F}}(l)\mathfrak{Z}_{\mathcal{F}}(k-l) = 0$ for all $l = 0, \dots, k$, from which it follows that $A_s = 0$. Therefore we have $S_{k,r,i,j} - (-1)^r S^\star_{k,r,i,j} = 0$, and the lemma follows. □

Computation of $S_{k,r,i,j}$. In this subsection, we shall compute $S_{k,r,i,j}$ (Lemma 3.9). The main ingredient of the computation is the following Ohno type relation, conjectured by Kaneko [2017] and established by Oyama [2018]:

Theorem 3.5 [Oyama 2018, Theorem 1.4]. *Let $\boldsymbol{k} = (k_1, \dots, k_r)$ be an index, and write its Hoffman dual as $\boldsymbol{k}^\vee = (k'_1, \dots, k'_{r'})$. Then for $m \in \mathbb{Z}_{\geq 0}$, we have*

$$\sum_{\substack{e_1+\cdots+e_r=m \\ e_1,\dots,e_r\geq 0}} \zeta_{\mathcal{F}}(k_1+e_1, \dots, k_r+e_r) = \sum_{\substack{e'_1+\cdots+e'_{r'}=m \\ e'_1,\dots,e'_r\geq 0}} \zeta_{\mathcal{F}}((k'_1+e'_1, \dots, k'_{r'}+e'_{r'})^\vee).$$

Lemma 3.6. *We have*

$$S_{k,r,i,j} = \sum_{\substack{e'_1+e'_2+e'_3=k' \\ e'_1,e'_2,e'_3\geq 0}} \zeta_{\mathcal{F}}((i+e'_1, i'+e'_2, i''+e'_3)^\vee).$$

Proof. Theorem 3.5 shows that if $i = j$, then

$$\begin{aligned} S_{k,r,i,j} &= \sum_{\substack{e_1+\cdots+e_r=k' \\ e_1,\dots,e_r\geq 0}} \zeta_{\mathcal{F}}(1+e_1, \dots, 1+e_{i-1}, 3+e_i, 1+e_{i+1}, \dots, 1+e_r) \\ &= \sum_{\substack{e'_1+e'_2+e'_3=k' \\ e'_1,e'_2,e'_3\geq 0}} \zeta_{\mathcal{F}}((i+e'_1, i'+e'_2, i''+e'_3)^\vee), \end{aligned}$$

and that if $i < j$, then

$$\begin{aligned} S_{k,r,i,j} &= \sum_{\substack{e_1+\cdots+e_r=k' \\ e_1,\dots,e_r\geq 0}} \zeta_{\mathcal{F}}\big(1+e_1, \dots, 1+e_{i-1}, 2+e_i, \\ &\qquad\qquad 1+e_{i+1}, \dots, 1+e_{j-1}, 2+e_j, 1+e_{j+1}, \dots, 1+e_r\big) \\ &= \sum_{\substack{e'_1+e'_2+e'_3=k' \\ e'_1,e'_2,e'_3\geq 0}} \zeta_{\mathcal{F}}((i+e'_1, i'+e'_2, i''+e'_3)^\vee). \end{aligned}$$

□

Lemma 3.7. *We have*

$$S_{k,r,i,j} = \frac{1}{2} \sum_{\substack{e'_1+e'_2+e'_3=k' \\ e'_1,e'_2,e'_3\geq 0}} \left(\begin{bmatrix} k \\ i+e'_1 \end{bmatrix} - \begin{bmatrix} k \\ i''+e'_3 \end{bmatrix} \right) \mathfrak{Z}_{\mathcal{F}}(k).$$

Proof. Using the same convention as in the statement of Lemma 2.9, we have

$$(i+e'_1, i'+e'_2, i''+e'_3)^\vee = (\{1\}^{i+e'_1-1}, 2, \{1\}^{i'+e'_2-2}, 2, \{1\}^{i''+e'_3-1}),$$

and so by Lemmas 2.9 and 3.6, we have

$$\begin{aligned} S_{k,r,i,j} &= \sum_{\substack{e_1'+e_2'+e_3'=k' \\ e_1',e_2',e_3'\geq 0}} \zeta_{\mathcal{F}}((i+e_1', i'+e_2', i''+e_3')^{\vee}) \\ &= \sum_{\substack{e_1'+e_2'+e_3'=k' \\ e_1',e_2',e_3'\geq 0}} \zeta_{\mathcal{F}}(\{1\}^{i+e_1'-1}, 2, \{1\}^{i'+e_2'-2}, 2, \{1\}^{i''+e_3'-1}) \\ &= \frac{1}{2} \sum_{\substack{e_1'+e_2'+e_3'=k' \\ e_1',e_2',e_3'\geq 0}} \left(\begin{bmatrix} k \\ i+e_1' \end{bmatrix} - \begin{bmatrix} k \\ i''+e_3' \end{bmatrix} \right) \mathfrak{Z}_{\mathcal{F}}(k). \end{aligned}$$ □

Lemma 3.8. *We have*

$$\sum_{\substack{e_1'+e_2'+e_3'=k' \\ e_1',e_2',e_3'\geq 0}} \begin{bmatrix} k \\ i+e_1' \end{bmatrix} = (k'+i+1)\left(\begin{bmatrix} k-1 \\ k'+i \end{bmatrix} - \begin{bmatrix} k-1 \\ i-1 \end{bmatrix} \right) + k\left(\begin{bmatrix} k-2 \\ k'+i-1 \end{bmatrix} - \begin{bmatrix} k-2 \\ i-2 \end{bmatrix} \right),$$

$$\sum_{\substack{e_1'+e_2'+e_3'=k' \\ e_1',e_2',e_3'\geq 0}} \begin{bmatrix} k \\ i''+e_3' \end{bmatrix} = (k'+i''+1)\left(\begin{bmatrix} k-1 \\ k'+i'' \end{bmatrix} - \begin{bmatrix} k-1 \\ i''-1 \end{bmatrix} \right) + k\left(\begin{bmatrix} k-2 \\ k'+i''-1 \end{bmatrix} - \begin{bmatrix} k-2 \\ i''-2 \end{bmatrix} \right).$$

Proof. By symmetry, we only need to show the first equality, which can be seen as follows:

$$\begin{aligned} &\sum_{\substack{e_1'+e_2'+e_3'=k' \\ e_1',e_2',e_3'\geq 0}} \begin{bmatrix} k \\ i+e_1' \end{bmatrix} \\ &= \sum_{e_1'=0}^{k'} (-1)^{i+e_1'} (k'-e_1'+1) \binom{k}{i+e_1'} \\ &= \sum_{e_1'=0}^{k'} (-1)^{i+e_1'} ((k'+i+1)-(i+e_1')) \binom{k}{i+e_1'} \\ &= (k'+i+1) \sum_{e_1'=0}^{k'} (-1)^{i+e_1'} \binom{k}{i+e_1'} - k \sum_{e_1'=0}^{k'} (-1)^{i+e_1'} \binom{k-1}{i+e_1'-1} \\ &= (k'+i+1) \sum_{e_1'=0}^{k'} \left((-1)^{i+e_1'} \binom{k-1}{i+e_1'} - (-1)^{i+e_1'-1} \binom{k-1}{i+e_1'-1} \right) \\ &\qquad + k \sum_{e_1'=0}^{k'} \left((-1)^{i+e_1'-1} \binom{k-2}{i+e_1'-1} - (-1)^{i+e_1'-2} \binom{k-2}{i+e_1'-2} \right) \end{aligned}$$

$$= (k'+i+1)\left((-1)^{k'+i}\binom{k-1}{k'+i} - (-1)^{i-1}\binom{k-1}{i-1}\right)$$
$$+ k\left((-1)^{k'+i-1}\binom{k-2}{k'+i-1} - (-1)^{i-2}\binom{k-2}{i-2}\right)$$
$$= (k'+i+1)\left(\begin{bmatrix}k-1\\k'+i\end{bmatrix} - \begin{bmatrix}k-1\\i-1\end{bmatrix}\right) + k\left(\begin{bmatrix}k-2\\k'+i-1\end{bmatrix} - \begin{bmatrix}k-2\\i-2\end{bmatrix}\right). \qquad \square$$

Lemma 3.9. *We have*

$$S_{k,r,i,j} = \tfrac{1}{2}N_{k,r,i,j}\mathfrak{Z}_{\mathcal{F}}(k),$$

where $N_{k,r,i,j}$ is an integer given by

$$N_{k,r,i,j} = (k'+i+1)\left(\begin{bmatrix}k-1\\k'+i\end{bmatrix} - \begin{bmatrix}k-1\\i-1\end{bmatrix}\right) - (k'+i''+1)\left(\begin{bmatrix}k-1\\k'+i''\end{bmatrix} - \begin{bmatrix}k-1\\i''-1\end{bmatrix}\right)$$
$$+ k\left(\begin{bmatrix}k-2\\k'+i-1\end{bmatrix} - \begin{bmatrix}k-2\\i-2\end{bmatrix} - \begin{bmatrix}k-2\\k'+i''-1\end{bmatrix} + \begin{bmatrix}k-2\\i''-2\end{bmatrix}\right).$$

Proof. Immediate from Lemmas 3.7 and 3.8. $\square$

Lemmas 3.4 and 3.9 complete the proof of our main theorem (Theorem 3.1).

References

[Hoffman 2015] M. E. Hoffman, "Quasi-symmetric functions and mod p multiple harmonic sums", *Kyushu J. Math.* **69**:2 (2015), 345–366. MR Zbl

[Jarossay 2014] D. Jarossay, "Double mélange des multizêtas finis et multizêtas symétrisés", *C. R. Math. Acad. Sci. Paris* **352**:10 (2014), 767–771. MR Zbl

[Kaneko 2017] M. Kaneko, "Finite multiple zeta values", pp. 175–190 in *Various aspects of multiple zeta values* (Kyoto, 2013), edited by K. Ihara, RIMS Kôkyûroku Bessatsu **B68**, Res. Inst. Math. Sci., Kyoto, 2017. In Japanese. MR Zbl

[Kaneko and Zagier $\geq$ 2019] M. Kaneko and D. Zagier, "Finite multiple zeta values", in preparation.

[Murahara 2016] H. Murahara, "A note on finite real multiple zeta values", *Kyushu J. Math.* **70**:1 (2016), 197–204. MR Zbl

[Oyama 2018] K. Oyama, "Ohno-type relation for finite multiple zeta values", *Kyushu J. Math.* **72**:2 (2018), 277–285. MR Zbl

[Saito 2017] S. Saito, "Numerical tables of finite multiple zeta values", pp. 191–208 in *Various aspects of multiple zeta values* (Kyoto, 2013), edited by K. Ihara, RIMS Kôkyûroku Bessatsu **B68**, Res. Inst. Math. Sci., Kyoto, 2017. MR Zbl

[Saito and Wakabayashi 2015] S. Saito and N. Wakabayashi, "Sum formula for finite multiple zeta values", *J. Math. Soc. Japan* **67**:3 (2015), 1069–1076. MR Zbl

[Zhao 2016] J. Zhao, *Multiple zeta functions, multiple polylogarithms and their special values*, Ser. Number Theory Appl. **12**, World Sci., Hackensack, NJ, 2016. MR Zbl

Received January 8, 2018. Revised October 10, 2018.

HIDEKI MURAHARA
NAKAMURA GAKUEN UNIVERSITY GRADUATE SCHOOL
BEFU, JONAN-KU
FUKUOKA
JAPAN
hmurahara@nakamura-u.ac.jp

SHINGO SAITO
FACULTY OF ARTS AND SCIENCE
KYUSHU UNIVERSITY
MOTOOKA, NISHI-KU
FUKUOKA
JAPAN
ssaito@artsci.kyushu-u.ac.jp

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.337

FROBENIUS–SCHUR INDICATORS FOR NEAR-GROUP AND HAAGERUP–IZUMI FUSION CATEGORIES

HENRY TUCKER

Dedicated to Susan Montgomery

Ng and Schauenburg generalized higher Frobenius–Schur indicators to pivotal fusion categories and showed that these indicators may be computed utilizing the modular data of the Drinfel'd center of the given category. We consider two classes of fusion categories generated by a single noninvertible simple object: near groups, those fusion categories with one noninvertible object, and Haagerup–Izumi categories, those with one noninvertible object for every invertible object. Examples of both types arise as representations of finite or quantum groups or as Jones standard invariants of finite-depth Murray–von Neumann subfactors. We utilize the computations of the tube algebras due to Izumi and to Evans and Gannon to obtain formulae for the Frobenius–Schur indicators of objects in both of these families.

1. Introduction

Fusion categories appear in a wide variety of mathematics and physics. Their objects have the properties of complex representations of finite groups; in particular, they are semisimple and have duals and tensor products. Important examples of fusion categories come from the representations of Drinfel'd–Jimbo quantum groups and Jones standard invariants of Murray–von Neumann subfactors. From the point of view of these examples fusion categories encode symmetry data in the quantum setting in the same way that finite groups do in the classical setting. Classification problems for these categories do not come without considerable difficulty; therefore, it is of great interest to find and understand categorical invariants.

The classical Frobenius–Schur indicator for finite groups was introduced in 1906. It determines if and how a given group representation is self-dual. This was generalized to the setting of semisimple Hopf algebras by Linchenko and Montgomery [2000] and further to the setting of quasi-Hopf algebras [Mason and

MSC2010: primary 18D10, 16T05; secondary 46L37.
Keywords: tensor category, fusion rules, Frobenius–Schur indicator, Drinfel'd center, modular data, Haagerup subfactor, Hopf algebras.

Ng 2005; Ng and Schauenburg 2008] and to pivotal tensor categories [Ng and Schauenburg 2007b].

The FS indicators are a *complete invariant* for the Tambara–Yamagami categories [Basak and Johnson 2015]. These are the fusion categories having exactly one noninvertible simple object ρ where $\mathrm{Hom}_{\mathcal{C}}(\rho \otimes \rho, \rho) = 0$. In the present paper we consider the *near-group categories*: those with exactly one noninvertible simple object ρ where $\dim_{\mathbb{C}}(\mathrm{Hom}_{\mathcal{C}}(\rho \otimes \rho, \rho)) = m$. (The Tambara–Yamagami categories are near groups with $m = 0$.) We provide the required background on this in Section 2.

Letting G be the group of invertible objects in our near-group category, we find in Section 3 that for the near-group categories with $m = |G| - 1$ the indicators are a complete invariant:

Corollary 3.3. *The near-group categories with $m = |G| - 1$ are completely distinguished by their Frobenius–Schur indicators.*

To make the computations here we utilize [Ng and Schauenburg 2007a, Theorem 4.1]: the Frobenius–Schur indicators of a spherical fusion category can be computed using the *ribbon structure* of the Drinfel'd center of the category. A complete list of near-group fusion categories in the case where $m = |G| \leq 13$ was found in [Evans and Gannon 2014]. In each of these examples the modular data for the Drinfel'd centers are given by quadratic forms. From this we get in Section 4:

Theorem 4.8. *In all known near-group categories with $m = |G|$ the noninvertible object has Frobenius–Schur indicators given by quadratic Gauss sums.*

This theorem provides new evidence for [Evans and Gannon 2014, Conjecture 2]: the *modular data* (matrix invariants from the braiding) of the centers of these near groups are always given by quadratic forms. The form of the indicators strongly suggests that these centers are formed from some "crossed product" construction for modular categories. See also the "pasting" of modular data developed in [Evans and Gannon 2011].

Finally, in Section 5, we observe a similar result which supports a similar conjecture for the *Haagerup–Izumi categories*, which are a related family of singly generated fusion categories having one noninvertible object for *each* invertible object:

Theorem 5.4. *All known Haagerup–Izumi categories have Frobenius–Schur indicators given by quadratic Gauss sums.*

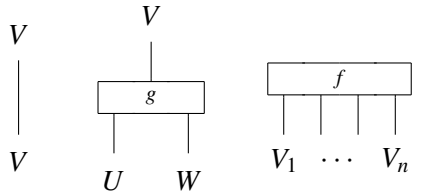
2. Categorical invariants

Tensor categories are abelian monoidal categories $(\mathcal{C}, \otimes, \mathbb{1})$ enriched over complex vector spaces; see [Etingof et al. 2015] or [Bakalov and Kirillov 2001] for the

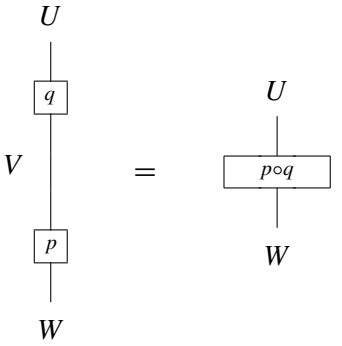
specifics of these definitions. The Mac Lane strictness theorem allows us the working assumption of *strictness*: the associativity natural isomorphism is the identity morphism for every triple of objects. Thus, we may use diagrammatic notation for the morphisms in these categories. Our notation is read from top to bottom, tensor products are given by side-by-side concatenation, and $\mathbb{1}$ is not written at all. For examples, the morphisms

$$\mathrm{id}_V : V \to V, \qquad g : V \to U \otimes W, \qquad f : \mathbb{1} \to V_1 \otimes \cdots \otimes V_n$$

are rendered in diagrammatic notation, respectively, as



Composition of morphisms is given by stacking; for example, given morphisms $p : V \to W$ and $q : U \to V$, their composition is given by



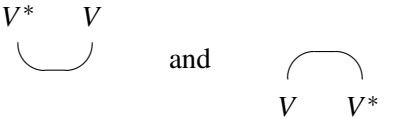
Categorifications of semisimple rings. Tensor categories should be thought of as a *categorification* of the notion of a unital algebra. The abelian and monoidal categorical structures are analogues of addition and multiplication, respectively. This point of view asks an obvious question:

What are the tensor categories that categorify a given ring?

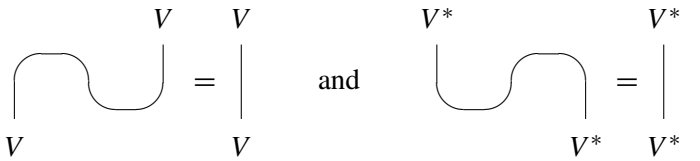
This question has produced several different interesting classification results for *semisimple* tensor categories: those where every object is a direct sum of some irreducible objects [Tambara and Yamagami 1998; Izumi 2001; Evans and Gannon 2014; 2017]. The set of (isomorphism classes) of the irreducible objects is denoted $\mathcal{I}rr(\mathcal{C})$.

Here we consider *fusion categories*: these are semisimple tensor categories $(\mathcal{C}, \otimes, \mathbb{1})$ that are additionally:

- *finitely* semisimple: $|\mathcal{I}rr(\mathcal{C})| < \infty$ and $\mathbb{1} \in \mathcal{I}rr(\mathcal{C})$.
- *rigid*: objects $V \in \mathcal{C}$ have duals $V^* \in \mathcal{C}$ with corresponding maps $\mathrm{ev}_V : V^* \otimes V \to \mathbb{1}$ and $\mathrm{db}_V : \mathbb{1} \to V \otimes V^*$. These are given, respectively, by the diagrams



satisfying the relations



These two requirements are meant to make the objects of $\mathcal{C}$ behave like group representations. Indeed, the tensor category $\mathrm{Rep}(G)$ of complex representations of a finite group G is the prototypical example of a fusion category: Maschke's theorem gives finite semisimplicity and the contragredient representation gives rigidity.

Now we make precise the notion of categorification. The *Grothendieck ring* $K_0(\mathcal{C})$ of a fusion category $\mathcal{C}$ is the $\mathbb{Z}$-based ring with basis $\mathcal{I}rr(\mathcal{C})$, multiplication given by the tensor product in $\mathcal{C}$, and addition given by the direct sum in $\mathcal{C}$; that is, $K_0(\mathcal{C})$ captures the ring structure of the category and forgets the morphisms. In the example of $\mathrm{Rep}(G)$ it is the character ring $R(G)$. We say that $\mathcal{C}$ *categorifies* a ring K if $K_0(\mathcal{C}) = K$.

The simplest class of based rings to consider are the group rings $\mathbb{Z}G$, which are categorified precisely by the *pointed* fusion categories. These are the categories Vec_G^ω of G-graded vector spaces where the associativity morphism for the tensor product of three irreducible objects is given by a 3-cocycle $\omega \in Z^3(G, \mathbb{C}^\times)$. These categories are classified to equivalence by the cohomology class of $[\omega] \in H^3(G, \mathbb{C}^\times)$. These facts are due to Mac Lane [Etingof et al. 2015, Proposition 4.10.3].

Here we will consider another level of complication, based rings with (only) one noninvertible object:

Definition 2.1. Let G be a finite group. A fusion category $\mathcal{C}$ is a *near group* if its Grothendieck ring is given by

$$K_0(\mathcal{C}) = \mathrm{NG}(G, m) := \mathbb{Z}[G \cup \{\rho\}]$$

where multiplication is given by the group law and, where $g \in G$,

$$\rho g = \rho = g\rho \quad \text{and} \quad \rho^2 = m\rho + \sum_{h \in G} h.$$

Remark 2.2 [Evans and Gannon 2014, Theorem 2(a)]. When G is abelian the multiplicity m is restricted to the values

- $m = |G| - 1$ or
- $m = k|G|$ for some $k \in \mathbb{N}$.

Consider the following important examples.

(1) The representation categories for the dihedral group of order 8 and the quaternion group of order 8 both categorify $\mathrm{NG}(\mathbb{Z}/(2) \times \mathbb{Z}/(2), 0)$. These are examples of *Tambara–Yamagami* categories, which are the near groups with $m = 0$ [Tambara and Yamagami 1998].

(2) $\mathrm{Rep}(S_3)$ and $\mathrm{Rep}(A_4)$ categorify $\mathrm{NG}(\mathbb{Z}/(2), 1)$ and $\mathrm{NG}(\mathbb{Z}/(3), 2)$, respectively. These are examples where $m = |G| - 1$.

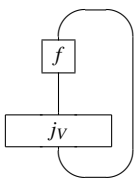
(3) The principal even sectors of the D_5 Murray–von Neumann subfactor also categorify $\mathrm{NG}(\mathbb{Z}/(2), 1)$.

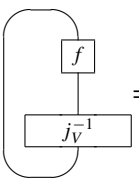
(4) $\mathrm{Rep}(\mathrm{AGL}_1(\mathbb{F}_q))$ categorifies $\mathrm{NG}(\mathbb{Z}/(q-1), q-2)$.

(5) The principal even sectors of the A_4, E_6, and Izumi–Xu subfactors categorify $\mathrm{NG}(\mathbb{Z}/(1), 1)$, $\mathrm{NG}(\mathbb{Z}/(2), 2)$, and $\mathrm{NG}(\mathbb{Z}/(3), 3)$, respectively.

Frobenius–Schur indicators. It is known that the Grothendieck ring and the associativity natural isomorphism completely determine a fusion category up to monoidal equivalence [Etingof et al. 2015, §§4.9–4.10]. The associativity data is encoded by the *6j symbols*, which are the matrix components of the linear maps induced by the associativity natural isomorphism on triples of simple objects. Directly classifying all 6j symbols having a given Grothendieck ring is difficult in general as it requires finding solutions to large systems of nonlinear equations.

The near groups are *spherical* fusion categories. These are the fusion categories equipped with a natural isomorphism of monoidal functors $j : \mathrm{Id}_{\mathcal{C}} \xrightarrow{\sim} (\,\cdot\,)^{**}$ (that is, a *pivotal structure*) whose associated left and right *quantum* (*or categorical*) *trace* functions agree for all objects $V \in \mathcal{I}rr(\mathcal{C})$ and morphisms $f \in \mathrm{Hom}_{\mathcal{C}}(V, V)$:

$\mathrm{qtr}^r(f) :=$ 

$V^* =\ V^*$ 

$=: \mathrm{qtr}^l(f).$

Note that these are complex numbers since we take V to be a simple object. Since the traces agree we are able to define a *quantum* (*or categorical*) *dimension* of objects $V \in \mathcal{C}$ by

$$\mathrm{qdim}(V) := \mathrm{qtr}(\mathrm{id}_V) = \mathrm{qtr}^r(\mathrm{id}_V) = \mathrm{qtr}^l(\mathrm{id}_V).$$

The name *spherical* is motivated by imagining the strings in the morphism diagrams to be inhabiting a *sphere* rather than a plane — this allows for the strands on either side of the quantum traces to rotate around, giving the equality pictured above in our planar diagrams.

For $\mathcal{C}$ a pivotal fusion category we can define a finer categorical invariant than the Grothendieck ring:

Definition 2.3 [Ng and Schauenburg 2007b]. For $V \in \mathcal{C}$ we define the *k-th Frobenius–Schur indicator* by the linear trace

$$\nu_k(V) = \mathrm{Tr}\left(E_V^{(k)} : \underbrace{V \ V \cdots V}_{n} \overset{f}{\mapsto} V \cdots V \xrightarrow{f} V \xrightarrow{j_V^{-1}} V \right)$$

with $E_V^{(k)}$ a linear endomorphism of finite-dimensional vector space $\mathrm{Hom}(\mathbb{1}, V^{\otimes n})$ taking $V^{\otimes n}$ to be n-fold tensor product of V with all parentheses to the right.

The Tambara–Yamagami categories are an example of a fusion category family where the Frobenius–Schur indicators are a finer invariant than the Grothendieck ring [Ng and Schauenburg 2008]. In [Basak and Johnson 2015] it was shown that the indicators are a *complete* invariant for the Tambara–Yamagami categories. That is, the monoidal equivalence classes of fusion categories associated to the ring $\mathrm{NG}(G, 0)$ are completely distinguished by their Frobenius–Schur indicators. We give this property a name:

Definition 2.4. A ring K exhibits *FS indicator rigidity* if its categorifications can all be distinguished by their Frobenius–Schur indicators.

The central question that motivates the present article is immediate:

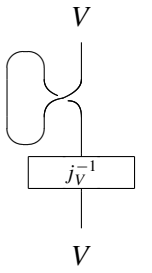
What rings have FS indicator rigidity?

We will see in Corollary 3.3 that the near-group rings $\mathrm{NG}(G, |G| - 1)$ exhibit this property.

Drinfel'd centers and modular data. The Drinfel'd center $\mathcal{Z}(\mathcal{C})$ of a spherical fusion category $\mathcal{C}$ is *modular* [Müger 2003, Proposition 5.10]; that is, it is again spherical with a *nondegenerate braiding* $c_{V,W} : V \otimes W \to V \otimes W$ which is given in diagrams by

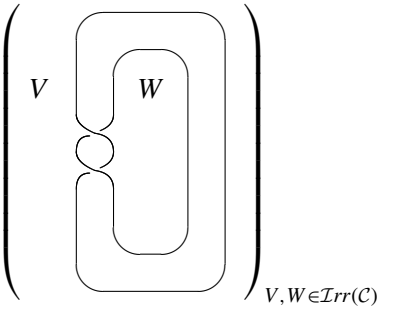
$$c_{V,W} = \begin{array}{c} V \ \ W \\ \diagup\!\diagdown \\ W \ \ V \end{array}$$

Combined with the spherical structure the braiding also endows the Drinfel'd center with a *ribbon structure*, that is, a natural isomorphism θ of the identity functor (satisfying some coherence axioms) given by

$\theta_V =$ 

Note that this is a scalar when V is a simple object.

Modular categories come with a projective representation of the modular group called *modular data*. The representation is defined by sending the generators $\mathfrak{s}, \mathfrak{t} \in SL_2(\mathbb{Z})$ to the *S- and T-matrices*

$S =$ 

$_{V,W \in \mathcal{I}rr(\mathcal{C})}$ and $T = \mathrm{Diag}(\theta_V)_{V \in \mathcal{I}rr(\mathcal{C})}$.

Most crucially, we can obtain the Frobenius–Schur indicators from the modular data of the Drinfel'd center:

Theorem 2.5 [Ng and Schauenburg 2007a, Theorem 4.1]. *Let $\mathcal{C}$ be a spherical fusion category, and let $F : \mathcal{Z}(\mathcal{C}) \to \mathcal{C}$ be the forgetful functor. Then*

$$\nu_k(X) = \frac{1}{\mathrm{qdim}(\mathcal{C})} \sum_{V \in \mathcal{I}rr(\mathcal{Z}(\mathcal{C}))} \theta_V^k \,\mathrm{qdim}(V) \dim_{\mathbb{C}}(\mathrm{Hom}_{\mathcal{C}}(F(V), X))$$

where θ_V are the entries of the T-matrix for $\mathcal{Z}(\mathcal{C})$.

Izumi's classification program. The classification parameters for the Tambara–Yamagami categories were obtained by direct solution of the equations resulting from the pentagon axiom for the associativity. This method is not feasible for more complicated categories.

Masaki Izumi was able to extend this classification by using a fundamental result due to Popa: every *unitary* (or C^*) fusion category tensor-generated by one object can be embedded in the category of sectors of the hyperfinite type-III

Murray–von Neumann subfactor R. Sectors are unitary equivalence classes of endomorphisms of R; the tensor product of sectors is composition. Note that the 6j symbols can be obtained from Izumi's classification data; see [Suzuki and Wakui 2002] for the near-group category $\mathcal{C}$ with $K_0(\mathcal{C}) = \mathrm{NG}(\mathbb{Z}/(3), 3)$ coming from the E_6 subfactor.

Izumi [2001] and Evans and Gannon [2014; 2017] have obtained the classification parameters for the near-group families used in the sequel via this program; hence, our fusion categories will be unitary. In particular, this means that a *canonical spherical structure* can be chosen such that the *quantum dimension and the Frobenius–Perron dimension agree*.

3. Frobenius–Schur indicators for near groups with $m = |G| - 1$

Let $\mathcal{C}$ be a fusion category such that $K_0(\mathcal{C}) = \mathrm{NG}(G, |G|-1)$. It is shown in [Evans and Gannon 2014, Proposition 2] that such a fusion category can only exist if $G \cong \mathbb{F}^\times_{|G|+1}$ is the multiplication group of a finite field. (So G is cyclic, and thus, $H^2(G, \mathbb{T}) = 1$.) Let $p = \mathrm{char}(\mathbb{F}_{|G|+1})$.

Consider again the category $\mathrm{Rep}(\mathrm{AGL}_1(\mathbb{F}_{|G|+1}))$. These provide the main examples of $m = |G| - 1$ near groups. In fact, by [Etingof et al. 2004, Corollary 7.4; Evans and Gannon 2014, Proposition 5], these are the *only* fusion categories with this Grothendieck ring unless $|G| = 1, 2, 3, 7$.

Indicators for $\mathcal{C} \simeq \mathbf{Rep}(\mathbf{AGL_1}(\mathbb{F}_q))$. We may use classical methods to determine the indicators for $\mathcal{C}$ that is tensor equivalent to the category of representations of an affine general linear group of degree 1 over the finite field $\mathbb{F}_q$. Recall that $\theta_k^G(h) = |\{g \in G \mid g^k = h\}|$.

Proposition 3.1. *Suppose $\mathcal{C}$ is such that $K_0(\mathcal{C}) = \mathrm{NG}(G, |G| - 1)$ and $|G| \neq 1, 2, 3, 7$. Then $\mathcal{C} \simeq_\otimes \mathrm{Rep}(\mathrm{AGL}_1(\mathbb{F}_{|G|+1}))$ and*

$$\nu_k(\rho) = \theta_k^G(e) - 1 + \delta_{\lfloor \frac{k}{p} \rfloor, \frac{k}{p}}.$$

Proof. Let $|G| + 1 = q$. Since $\mathrm{AGL}_1(\mathbb{F}_q) \cong \mathbb{F}_q^+ \rtimes GL_1(\mathbb{F}_q) \cong \mathbb{F}_q^+ \rtimes \mathbb{F}_q^\times$ we may use Serre's method of little groups [1977, §8.2, Proposition 25] to see that the character ρ for the irreducible representation with degree > 1 is given by

$$\rho(a, b) = \frac{\delta_{1,b}}{q} \sum_{(x,y) \in \mathrm{AGL}_1(\mathbb{F}_q)} \eta(y^{-1}a)$$

for any *nontrivial* linear character $\eta \in \widehat{\mathbb{F}_q^+}$.

Now we may apply the classical formula for $\nu_k(\rho)$ [Isaacs 1976, Lemma 4.4]:

$$\begin{aligned}\nu_k(\rho) &= \frac{1}{q(q-1)} \sum_{(a,b)\in\mathbb{F}_q\rtimes\mathbb{F}_q^\times} \rho((a,b)^k) \\ &= \frac{1}{q(q-1)} \sum_{(a,b),\, b^k=1} \rho((1+b+b^2+\cdots+b^{k-1})a, 1) \\ &= \frac{1}{q(q-1)} \left(\sum_{(a,b),\, b^k=1,\, b\neq 1} \rho(0,1) + \sum_{n\in\mathbb{F}_q} \rho(kn,1) \right).\end{aligned}$$

Since ρ is a degree-$(q-1)$ character, the left-hand sum in the last expression above is equal to $q(q-1)(\theta_k^{\mathbb{F}_q^\times}(1)-1)$. The right-hand sum in the same expression is equal to

$$\sum_{n\in\mathbb{F}_q} \sum_{b\in\mathbb{F}_q^\times} \eta(b^{-1}kn) = \begin{cases} q(q-1) & \text{if } p \mid k, \\ 0 & \text{if } p \nmid k. \end{cases}$$

The $p \mid k$ case is clear since then $\eta(b^{-1}kn)$ is identically 1. On the other hand, $\eta(b^{-1}kn) = b\cdot\eta(kn)$ under the transpose of the left regular action of $\mathbb{F}_q^\times \cong GL_1(\mathbb{F}_q)$ on $\widehat{\mathbb{F}_q} \cong \mathbb{F}_q$. Since $(p,k)=1$ we have that

$$\sum_{n\in\mathbb{F}_q} b\cdot\eta(kn) = \sum_{n\in\mathbb{F}_q} b\cdot\eta(n),$$

and since the action is faithful by definition, we know that $b\cdot\eta$ is not the trivial representation for any $b\in\mathbb{F}_q^\times$. Hence, by orthogonality of characters the sum is 0. The formula is now clear since the given Kronecker delta is 1 if $p \mid k$ and is 0 otherwise. □

Indicators in general from modular data of $\mathcal{Z}(\mathcal{C})$. For $|G| = 1, 3, 7$ there is 1 additional monoidal equivalence class, and for $|G| = 2$ there are 2 additional monoidal equivalence classes. The modular data for Drinfel'd centers of unitary $m = |G| - 1$ near groups was computed in [Evans and Gannon 2014, Theorem 5]. We will appeal to Theorem 2.5 to compute the indicators for a general unitary $m = |G| - 1$ near group.

Let $\epsilon \in \widehat{G}$ be the trivial character, and let $\mathbb{F}^+_{|G|+1}$ be the additive group of the finite field. Excluding the case where $|G| = 7$ and $s = -1$ we have the following data for the center $\mathcal{Z}(\mathcal{C})$:

$X \in \mathcal{I}rr(\mathcal{Z}(\mathcal{C}))$		$F(X)$	$c_{X,\cdot}$ given by	θ_X
A_g	$(g \in G)$	g	1	1
Σ		$\bigoplus_{x\in G} x$	1	1
B_g^ω	$(g \in G)$	$\rho + g$	$\omega \in \widehat{G} \setminus \{\epsilon\}$	$\overline{\omega(g)}$
C^ψ	$(\psi \in \widehat{\mathbb{F}^+_{\lvert G\rvert+1}})$	ρ	$\psi \in \widehat{\mathbb{F}^+_{\lvert G\rvert+1}}$	$\overline{\zeta_1\psi(1)}$

where the half-braiding for C^ψ on occurrences of ρ in objects of $\mathcal{Z}(\mathcal{C})$ is a morphism

$$e_{C^\psi}(\rho) \in \mathrm{Hom}_{\mathcal{C}}(\rho \otimes \rho, \rho \otimes \rho) \cong \mathbb{C}^{|G|} \oplus M_m(\mathbb{C})$$

given by

$$e_{\mathcal{C}^\psi}(\rho) = \zeta_1\psi(1)\left(\bigoplus_{k\in G}(-1)^{mk}\,\mathrm{Id}_k\right) \oplus [\zeta_\gamma(\psi \circ \sigma)(\gamma)\delta_{\sigma^2(\gamma)^*,\mu}\,\mathrm{Id}_\rho]_{\gamma,\mu}.$$

For the case where $|G| = 7$ and $s = -1$ we have:

$X \in \mathcal{I}rr(\mathcal{Z}(\mathcal{C}))$		$F(X)$	$c_{X,\cdot}$ given by	θ_X
A_g	$(g \in G)$	g	1	1
Σ		$\bigoplus_{x\in G} x$	1	1
B_g^ω	$(g \in G,\ \omega \in \widehat{G} \setminus \{\epsilon\})$	$\rho + g$	$\omega \in \widehat{G} \setminus \{\epsilon\}$	$\overline{\omega(g)}$
E_1		$\rho + \rho$	1	i
E_2		$\rho + \rho$	1	$-i$

With the preceding data in hand we may now apply Theorem 2.5 to see:

Theorem 3.2. *Suppose that $\mathcal{C}$ is a unitary fusion category such that $K_0(\mathcal{C}) = \mathrm{NG}(G, |G| - 1)$. Then the indicators for the noninvertible object ρ are given by*:

(1) *If $|G| \neq 7$ or $s = 1$, then*

$$\nu_k(\rho) = (\theta_k^G(e) - 1) + \overline{\zeta_1}^k \delta_{\lfloor \frac{k}{p} \rfloor, \frac{k}{p}}.$$

(2) *If $|G| = 7$ and $s = -1$, then*

$$\nu_k(\rho) = (\theta_k^G(e) - 1) + (-1)^{k/2}\delta_{\lfloor \frac{k}{2} \rfloor, \frac{k}{2}}.$$

Proof. (1) Suppose $|G| \neq 7$ or $s = 1$. Then we have

$$\begin{aligned}\nu_k(\rho) &= \frac{1}{\mathrm{qdim}(\mathcal{C})}\left(\sum_{\substack{g\in G\\ \omega\in\widehat{G}\setminus\{\epsilon\}}} \theta^k_{B_g^\omega}\,\mathrm{qdim}(B_g^\omega) + \sum_{\psi\in\widehat{\mathbb{F}^+_{|G|+1}}} \theta^k_{C^\theta}\,\mathrm{qdim}(C^\psi)\right)\\ &= \frac{1}{\mathrm{qdim}(\mathcal{C})}\left((|G|+1)\sum_{\substack{g\in G\\ \omega\in\widehat{G}\setminus\{\epsilon\}}} \overline{\omega(g)}^k + |G|\overline{\zeta_1}^k \sum_{\psi\in\widehat{\mathbb{F}^+_{|G|+1}}} \overline{\psi(1)}^k\right).\end{aligned}$$

Consider the first summand. Since G is abelian we may choose an isomorphism $h \mapsto \chi_h$ from $G \to \widehat{G}$. Then we have

$$\begin{aligned}\sum_{\substack{g\in G\\ \omega\in\widehat{G}\setminus\{\epsilon\}}} \overline{\omega(g)}^k &= \left(\sum_{g\in G}\sum_{\omega\in\widehat{G}} \overline{\omega(g)}^k\right) - \left(\sum_{g\in G}\overline{\epsilon(g)}^k\right)\\ &= \left(\sum_{g\in G}\sum_{h\in G}\overline{\chi_h(g)}^k\right) - |G|\\ &= \left(|G|\sum_{h\in G}\overline{\nu_k(\chi_h)}\right) - |G|\\ &= |G|(\theta_k^G(e)-1).\end{aligned}$$

Consider the second summand. Since $\mathbb{F}_{n+1}^+$ is the additive group of a finite field we have that $n+1 = p^l$ for some prime p and positive integer l and that $\mathbb{F}_{n+1}^+ \cong (\mathbb{Z}_p)^l$ as groups. Under this identification the multiplicative unit $1 \in \mathbb{F}_{n+1}^+$ is a direct sum of generators of the copies of $\mathbb{Z}_p$:

$$\begin{aligned}\sum_{\psi\in\widehat{\mathbb{F}_{|G|+1}^+}} \overline{\psi(1)}^k = \sum_{\psi\in\widehat{\mathbb{F}_{|G|+1}^+}} \overline{\psi(k1)} &= \begin{cases}0 & \text{if } k1\neq 0,\\ p^l & \text{if } k1 = 0\end{cases}\\ &= \begin{cases}0 & \text{if } p\nmid k,\\ |G|+1 & \text{if } p\mid k\end{cases}\\ &= (|G|+1)\delta_{\lfloor \frac{k}{p}\rfloor,\frac{k}{p}}.\end{aligned}$$

(2) Now suppose that $|G| = 7$ and $s = -1$. Then

$$\begin{aligned}\nu_k(\rho) &= \frac{1}{\operatorname{qdim}(\mathcal{C})}\left(\sum_{\substack{g\in G\\ \omega\in\widehat{G}\setminus\{\epsilon\}}} \theta_{B_g^\omega}^k \operatorname{qdim}(B_g^\omega) + 2\sum_{i=1}^{2}\theta_{E_t}^k \operatorname{qdim}(E_t)\right)\\ &= \theta_k^G(e) - 1 + \frac{4|G|i^k(1+(-1)^k)}{|G|+|G|^2}\\ &= \theta_k^G(e) - 1 + \frac{i^k(1+(-1)^k)}{2}\\ &= \theta_k^G(e) - 1 + (-1)^{k/2}\delta_{\lfloor \frac{k}{2}\rfloor,\frac{k}{2}}.\end{aligned}$$

□

Corollary 3.3. *The near-group fusion ring* $\mathrm{NG}(G, |G|-1)$ *exhibits Frobenius–Schur indicator rigidity.*

Proof. The statement is vacuous in all but the cases where $|G| = 1, 2, 3, 7$. We shall consider them now.

If $|G| = 1, 3, 7$, then there is one additional tensor equivalence class corresponding to $s = -1$. By [Evans and Gannon 2014, p. 41] if $|G| + 1$ is even (i.e., a power of 2), then $\zeta_1^2 = s$; hence, $\nu_2(\rho) = s$ in each of these three cases.

If $|G| = 2$, then $s = 1$ but instead $b = \mu$ where μ is some third root of unity. The two nontrivial possibilities for μ correspond to the two additional tensor equivalence classes for this type. By [Evans and Gannon 2014, p. 42] if $\mu = \exp\left(\pm\frac{2\pi i}{3}\right)$, then $\zeta_1 = \exp\left(\mp\frac{2\pi i}{3}\right)$; hence, $\nu_3(\rho) = \mu$. □

4. Frobenius–Schur indicators for near groups with $m = |G|$

For the rest of this article the group operation in G will be written *additively*. This will be a more convenient notation for working with bilinear and quadratic forms.

Metric groups and the Fourier transform. Shimizu observed that the Fourier transform for finite groups appears when computing Frobenius–Shur indicators for fusion categories [Shimizu 2011]. A finite abelian group G is isomorphic to its linear dual $\widehat{G}$ via a nondegenerate symmetric *bicharacter* $\langle \cdot, \cdot \rangle$ with the identification

$$G \to \widehat{G},$$
$$g \mapsto \langle g, \cdot \rangle.$$

Symmetric bicharacters $\langle \cdot, \cdot \rangle : G \times G \to \mathbb{T}$ are in one-to-one correspondence with bilinear forms $\beta : G \times G \to \mathbb{Q}/\mathbb{Z}$ via the exponential

$$\langle g, h \rangle = e^{2\pi i \beta(g,h)}.$$

A *quadratic form* is a function $q : G \to \mathbb{Q}/\mathbb{Z}$ with $q(-g) = q(g)$ such that

$$\partial q(g, h) := q(g) + q(h) - q(gh)$$

is a symmetric bilinear form. A pair (G, q) is called a *premetric group*. If the bilinear form ∂q is nondegenerate, then it is called a *metric group*.

Remark 4.1. If $|G|$ is odd, then the correspondence between quadratic forms and bilinear forms given by ∂ is one-to-one. If $|G|$ is even, then the correspondence is $|G/2G|$-to-one.

Now, using the bicharacter $\langle \cdot, \cdot \rangle$ we can define the *Fourier transform* for complex function $f : G \to \mathbb{C}$ on finite abelian groups:

$$\hat{f}(g) = \frac{1}{\sqrt{|G|}} \sum_{h \in G} \overline{\langle g, h \rangle} f(h).$$

The Fourier transform of the exponent of a quadratic form q at the unit element of the group defines a very important invariant of premetric groups:

Definition 4.2. Let (G, q) be a premetric group. Then the Fourier transform of $e^{2\pi iq}$ at $0 \in G$ defines the *Gauss sum*:

$$\Theta(G, q) = \frac{1}{\sqrt{|G|}} \sum_{g \in G} e^{2\pi i q(g)}.$$

The Gauss sum is multiplicative over the (obviously defined) orthogonal direct product of metric groups:

$$\Theta(G \perp G', q + q') = \Theta(G, q)\Theta(G', q').$$

(This identity is for *metric* groups; hence, the quadratic forms must all be nondegenerate.)

Izumi's classification of $m = |G|$ near groups. Izumi completely classified unitary near-group fusion categories with $m = |G|$ and where $H^2(G, \mathbb{C}^\times) = 1$ in [Izumi 2000; 2001; 2017].

Theorem 4.3 [Izumi 2001, Theorem 5.3]. *Unitary fusion categories $\mathcal{C}$ such that $K_0(\mathcal{C}) = \mathrm{NG}(G, |G|)$ and $H^2(G, \mathbb{C}^\times) = 1$ are classified up to monoidal equivalence by the group G, a metric group structure $\langle \cdot, \cdot \rangle$ on G, and the following complex parameters*:

(1) *$a : G \to \mathbb{T}$ such that $a(g) = e^{2\pi i q(g)}$ for a quadratic form q with $\langle g, h \rangle = e^{2\pi i \partial q(g,h)}$, i.e.,*

$$a(0) = 1, \qquad a(g) = a(-g), \qquad \frac{a(g+h)}{a(g)a(h)} = \langle g, h \rangle.$$

(2) *$b : G \to \mathbb{C}$ and $c \in \mathbb{T}$ such that*

$$\Theta(G, q) = \hat{a}(0) = \frac{1}{c^3},$$
$$b(g) = \overline{a(g)b(-g)},$$
$$\hat{b}(0) = \frac{-c}{\mathrm{qdim}(\rho)}, \qquad \hat{b}(g) = c\overline{b(g)}, \qquad |\hat{b}(g)|^2 = \frac{1}{|G|} - \frac{\delta_{g,0}}{\mathrm{qdim}(X)},$$
$$\sum_{x \in G} b(x+g)b(x+h)\overline{b(x)} = \langle g, h \rangle b(g)b(h) - \frac{c}{\mathrm{qdim}(\rho)\sqrt{|G|}}.$$

Two such fusion categories $\mathcal{NG}(G_1, \langle \cdot, \cdot \rangle_1, a_1, b_1, c_1), \mathcal{NG}(G_2, \langle \cdot, \cdot \rangle_2, a_2, b_2, c_2)$ are monoidally equivalent if and only if

$$c_1 = c_2$$

and there is an isomorphism of metric groups $\phi : (G_1, \langle \cdot, \cdot \rangle_1) \to (G_2, \langle \cdot, \cdot \rangle_2)$ such that

$$a_2 = a_1 \circ \phi \quad \text{and} \quad b_2 = b_1 \circ \phi.$$

Remark 4.4. If the requirement that $H^2(G, \mathbb{C}^\times) = 1$ is relaxed, then a solution of Izumi's equations in Theorem 4.3 is sufficient to produce a near-group category with the required K_0 ring, but *not* necessary.

Indicators from modular data of $\mathcal{Z}(\mathcal{C})$. Izumi found the simple objects of $\mathcal{Z}(\mathcal{C})$ along with their twists and half-braidings in [Izumi 2001, Theorem 6.8], which is given as follows, where $<$ is a chosen order on G:

$X \in \mathcal{I}rr(\mathcal{Z}(\mathcal{C}))$		$F(X)$	$c_{X,\cdot}$ given by	θ_X
A_g	$(g \in G)$	g	1	$\langle g, g\rangle$
B_g	$(g \in G)$	$\rho + g$	1	$\langle g, g\rangle$
$C_{g,h}$	$(g < h \in G)$	$\rho + g + h$	1	$\langle g, h\rangle$
E_j	for $j = 1, \ldots, \frac{1}{2}\lvert G\rvert(\lvert G\rvert + 3)$	ρ		ω_j

The $\omega_j \in \mu_\infty \subseteq \mathbb{T}$ are solutions to the system of equations (6.18)-(6.20) in [Izumi 2001, §6] parametrized by $g \in G$ with coefficients given by the complex values $a(g), b(g), c \in \mathbb{C}$.

Proposition 4.5. *Suppose $\mathcal{C}$ is unitary fusion category with Grothendieck ring $K_0(\mathcal{C}) = \mathrm{NG}(G, |G|)$ and noninvertible object ρ. Let q be a quadratic form such that $\langle g, h\rangle = e^{2\pi i \partial q(g,h)}$. Then the indicators for ρ are given by*

$$\nu_k(\rho) = \tfrac{1}{2}\theta_k^G(e) + \frac{\mathrm{qdim}(\rho)}{\mathrm{qdim}(\mathcal{C})}\left(\frac{\sqrt{|G|}}{2}\Theta(G, 2kq) + \sum_{j=1}^{|G|(|G|+3)/2} \omega_j^k\right).$$

Proof. Let $d_\rho := \mathrm{qdim}(\rho)$, and let $<$ be an arbitrary ordering on the finite group G. Again applying Theorem 2.5 we have

$$\begin{aligned}\nu_k(\rho) &= \frac{1}{\mathrm{qdim}(\mathcal{C})}\left((1+d_\rho)\sum_{g\in G}\theta_{B_g}^k + (2+d_\rho)\sum_{\substack{g,h\in G\\ g<h}}\theta_{C_{g,h}}^k + d_\rho\sum_{j=1}^{|G|(|G|+3)/2}\theta_{E_j}^k\right)\\ &= \frac{1}{\mathrm{qdim}(\mathcal{C})}\left(\frac{d_\rho}{2}\sum_{g\in G}\langle g, g\rangle^k + \frac{2+d_\rho}{2}\sum_{g,h\in G}\langle g, h\rangle^k + d_\rho\sum_j \omega_j^k\right)\end{aligned}$$

where the second equality is due to the symmetry of $\langle\cdot,\cdot\rangle$.

Now we consider the middle sum:

$$\sum_{g,h\in G}\langle g, h\rangle^k = \sum_{g\in G}\sum_{h\in G}\langle g, h^k\rangle = |G|\sum_{g\in G}\nu_k^{\text{groups}}(\langle g, \cdot\,\rangle) = |G|\theta_k^G(e).$$

The second equality is by definition of the Frobenius–Schur indicator for finite groups (denoted ν_k^{groups}) [Isaacs 1976, (4.4)] and the third equality is by [Isaacs 1976, p. 49].

Now let q be a quadratic form such that $\langle g, h\rangle = e^{2\pi i \partial q(g,h)}$, and consider the first sum:

$$\sum_{g\in G}\langle g, g\rangle^k = \sum_{g\in G} e^{2\pi i(2kq(g))} = \sqrt{|G|}\Theta(G, 2kq).$$

Hence, the formula is now clear. □

Modular data for pointed modular categories. Recall that any *pointed* fusion category is equivalent to Vec_G^ω for some $[\omega] \in H^3(G, \mathbb{C}^\times)$. Now we consider pointed *modular* categories. Since modular categories are also fusion categories they will be equivalent as fusion categories to Vec_G^ω with G abelian. The braiding induces a quadratic form $c_{g,g} = e^{2\pi i q(g)}$, which gives G the structure of a metric group. Then these categories are classified under *braided* equivalence up to isomorphism of *premetric* groups. Note that in the case of *odd-order* groups if ω admits a braiding it will be unique; the notational convention $\mathrm{Vec}_G^{(\omega,c)}$ includes the braiding c. (This is because the Eilenburg–Mac Lane abelian cohomology $H^3_{\mathrm{ab}}(G, \mathbb{C}^\times)$ is isomorphic to the group of quadratic forms on G [Eilenberg and Mac Lane 1953; 1954]. See [Etingof et al. 2015, §8.4] for an outline of the proof in a more modern context.)

We now give the modular data for a pointed modular category. Define the bicharacter $\langle g, h\rangle_q := e^{2\pi i \partial q(g,h)}$. The modular data are given by the *Weil representation* associated to the premetric group (G, q):

$$S = S^q := \frac{1}{\sqrt{|G|}}(\langle g, h\rangle_q)_{g,h\in G}, \qquad T = T^q := (\delta_{g,h} e^{2\pi i q(g)})_{g,h\in G}.$$

Indicators when $|G|$ is odd. When $|G|$ is odd we have a one-to-one between quadratic forms and bilinear forms given by the map $q \mapsto \partial q$. Let q be the quadratic form on G such that $\langle g, h\rangle = e^{2\pi i \partial q(g,h)}$. Then we define

$$\mathcal{NG}(G, q, b, c) := \mathcal{NG}(G, \langle\cdot,\cdot\rangle_q, e^{2\pi i q}, b, c),$$

the corresponding near-group fusion category via the notation from Theorem 4.3.

Conjecture 4.6 [Evans and Gannon 2014, Conjecture 2]. *Suppose $|G|$ is odd. Then there exists a metric group (G', q') of order $|G|+4$ such that*:

(1) *Simple objects E_j in the subsection starting on page 350 are indexed by $g \in G$ and $x \in G' \setminus \{e\}$ where $E_{g,x} = E_{g,x^{-1}}$ and*

$$\theta_{E_{g,x}} = \langle g, g\rangle e^{2\pi i \partial q'(x)}.$$

(2) *The modular data are given by the Kronecker product of the Weil representation for (G, q) with another pair of modular data (S', T') for a rank $|G|+3$ modular category*:

$$S^{q,q'} := S^q \otimes S', \qquad T^{q,q'} := T^q \otimes T'$$

where we have

$$T' = \mathrm{Diag}(1, 1, \langle g, g\rangle_q, \langle x, x\rangle_{q'})_{g\in G, x\in G'}.$$

(See [Evans and Gannon 2014, Proposition 7] for the definition of S'.)

Remark 4.7. Evans and Gannon [2014] show that the conjecture is true for near groups with $|G| \leq 13$ odd.

Theorem 4.8. *Suppose a unitary fusion category $\mathcal{C}$ with $K_0(\mathcal{C}) = \mathrm{NG}(G, |G|)$ and $|G|$ odd satisfies Conjecture 4.6. Then*

$$\nu_k(\rho) = \tfrac{1}{2}\theta_k^G(e) + \tfrac{1}{2}\Theta(G, 2kq)\Theta(G', 2kq').$$

Proof. Let $N = (|G'| - 1)/2$, and enumerate G' as

$$G' = \{e, x_1, \ldots, x_N, x_1^{-1}, \ldots, x_N^{-1}\}.$$

Let $d_\rho := \mathrm{qdim}(\rho)$ and $\langle x, y\rangle_{q'} := e^{2\pi i \partial q(x,y)}$. Starting with Proposition 4.5 we have

$$\begin{aligned}
\nu_k(\rho) =& \tfrac{1}{2}\theta_k^G(e) + \frac{d_\rho}{\mathrm{qdim}(\mathcal{C})}\left(\frac{\sqrt{|G|}}{2}\Theta(G, 2kq) + \sum_{\substack{g\in G\\ 1\leq i\leq N}} \langle g, g\rangle_q^k \langle x_i, x_i\rangle_{q'}^k\right)\\
=& \tfrac{1}{2}\theta_k^G(e) + \frac{d_\rho\sqrt{|G|}\Theta(G, 2kq)}{|G|(2+d_\rho)}\left(\tfrac{1}{2} + \sum_{1\leq i\leq N} \langle x_i, x_i\rangle_{q'}^k\right)\\
=& \tfrac{1}{2}\theta_k^G(e) + \frac{d_\rho\sqrt{|G|}\Theta(G, 2kq)}{|G|(2+d_\rho)}\left(\tfrac{1}{2} + \tfrac{1}{2}(\Theta(G', 2kq')\sqrt{|G|+4} - 1)\right)\\
=& \tfrac{1}{2}\theta_k^G(e) + \frac{d_\rho\sqrt{|G|}\sqrt{|G|+4}}{2|G|(2+d_\rho)}\Theta(G, 2kq)\Theta(G, 2kq'),
\end{aligned}$$

and using the fact that $d_\rho^2 = |G| + |G|d_\rho$ we have

$$\frac{d_\rho\sqrt{|G|}\sqrt{|G|+4}}{2|G|(2+d_\rho)} = \frac{d_\rho(2d_\rho - |G|)}{2|G|(2+d+\rho)} = \frac{1}{2}$$

and then the formula for $\nu_k(\rho)$ is clear. □

As a corollary to the preceding theorem we obtain an easy and more natural proof of [Evans and Gannon 2014, Proposition 7(b)]:

Corollary 4.9. *The matrices $(S^{q,q'}, T^{q,q'})$ are modular data for a near-group center only if $\Theta(G, 2q)\Theta(G', 2q') = -1$.*

Proof. Suppose a near-group category $\mathcal{C} = \mathcal{NG}(G, \langle\cdot,\cdot\rangle_q, e^{2\pi i q}, b, c)$ has a center $\mathcal{Z}(\mathcal{C})$ with modular data given by $(S^{q,q'}, T^{q,q'})$. Since the simple object ρ cannot contain a copy of $\mathbb{1}$ we know that $\nu_1(\rho) = 0$. Therefore, by Theorem 4.8, we must have $\Theta(G, 2q)\Theta(G', 2q') = -1$. □

Let $\left(\frac{p}{q}\right)$ be the Jacobi symbol.

Corollary 4.10. *For $|G|$ odd and k such that* $\gcd(k, |G| \cdot |G'|) = 1$ *we have*

$$\nu_k(\rho) = \frac{1}{2}\left(1 - \left(\frac{k}{|G| \cdot |G'|}\right)\right).$$

Proof. Using the decomposition into irreducible metric groups given in [Wall 1963] it is easy to see that

$$\Theta(G, kq)\Theta(G', kq') = \left(\frac{k}{|G| \cdot |G'|}\right)\Theta(G, q)\Theta(G', q').$$

See also [Basak and Johnson 2015, §3 and Lemma 3.2]. □

Corollary 4.11. $\Theta(G', q') = -c^3$ *where (G', q') is the metric group associated to the center of $\mathcal{NG}(G, q, b, c)$.*

Proof. We've seen in Theorem 4.3 that $\Theta(G, q) = \frac{1}{c^3}$; hence, the above follows by the Corollary 4.9. □

The complete list of near-group categories with $K_0(\mathcal{C}) = \mathrm{NG}(G, |G|)$ for odd $|G| \leq 13$ was obtained in [Evans and Gannon 2014, Proposition 6] by finding solutions to Izumi's equations in Theorem 4.3. They also used Izumi's methods from [Izumi 2001, §6; 2017] to produce the modular data of their Drinfel'd centers; see [Evans and Gannon 2014, §§4.3–4.4 and Table 2]. Collected below are the modular data they found along with the Frobenius–Schur indicators of ρ for each of these categories. Since $|G|$ is odd, let q be the *unique* quadratic form associated to the bicharacter $\langle \cdot, \cdot \rangle$ from the classification parameters.

The data uses the following notation:

- *Column 1.* $\mathcal{C} = \mathcal{NG}(G, q, b, c)$) with $|G|$ odd as in the above notation. (For clearer presentation, the parameters b and c will be given only if they are necessary to establish in-equivalence.)
- *Column 2.* (G', q') is the metric group from the modular data of $\mathcal{Z}(\mathcal{C})$ from Conjecture 4.6. Recall $|G'| = |G| + 4$.
- $\zeta_k = \exp\left(\frac{2\pi i}{k}\right) \in \mathbb{T}$ primitive k-th root of unity.

$\lvert G\rvert = 3$	(G', q')	$\nu_3(\rho)$	$\nu_7(\rho)$
$\mathcal{NG}(\mathbb{Z}/(3), \frac{1}{3}g^2, \cdot, \cdot)$	$(\mathbb{Z}/(7), \frac{1}{7}g^2)$	$\frac{3+i\sqrt{3}}{2}$	$\frac{1+i\sqrt{7}}{2}$
$\mathcal{NG}(\mathbb{Z}/(3), -\frac{1}{3}g^2, \cdot, \cdot)$	$(\mathbb{Z}/(7), -\frac{1}{7}g^2)$	$\frac{3-i\sqrt{3}}{2}$	$\frac{1-i\sqrt{7}}{2}$

$\lvert G\rvert = 5$	(G', q')	$\nu_3(\rho)$	$\nu_5(\rho)$	$\nu_9(\rho)$
$\mathcal{NG}(\mathbb{Z}/(5), \frac{2}{5}g^2, \cdot, \zeta_3)$	$(\mathbb{Z}/(9), \frac{2}{9}g^2)$	$1+\overline{\zeta_3}$	$\frac{5+\sqrt{5}}{2}$	-1
$\mathcal{NG}(\mathbb{Z}/(5), \frac{2}{5}g^2, \cdot, \overline{\zeta_3})$	$(\mathbb{Z}/(9), -\frac{2}{9}g^2)$	$1+\zeta_3$	$\frac{5+\sqrt{5}}{2}$	-1
$\mathcal{NG}(\mathbb{Z}/(5), \frac{1}{5}g^2, \cdot, 1)$	$((\mathbb{Z}/(3))^2, \frac{1}{3}(g^2+h^2))$	-1	$\frac{5-\sqrt{5}}{2}$	2

$\lvert G\rvert = 7$	(G', q')	$\nu_7(\rho)$	$\nu_{11}(\rho)$
$\mathcal{NG}(\mathbb{Z}/(7), \frac{1}{7}g^2, \cdot, \cdot)$	$(\mathbb{Z}/(11), -\frac{2}{11}g^2)$	$\frac{7-i\sqrt{7}}{2}$	$\frac{1+i\sqrt{11}}{2}$
$\mathcal{NG}(\mathbb{Z}/(7), -\frac{1}{7}g^2, \cdot, \cdot)$	$(\mathbb{Z}/(11), \frac{2}{11}g^2)$	$\frac{7+i\sqrt{7}}{2}$	$\frac{1-i\sqrt{11}}{2}$

$\lvert G\rvert = 9$	(G', q')	$\nu_3(\rho)$	$\nu_9(\rho)$	$\nu_{13}(\rho)$
$\mathcal{NG}(\mathbb{Z}/(9), \frac{1}{9}g^2, \cdot, \cdot)$	$(\mathbb{Z}/(13), -\frac{2}{13}g^2)$	$1-\zeta_3$	3	$\frac{1+\sqrt{13}}{2}$
$\mathcal{NG}(\mathbb{Z}/(9), -\frac{1}{9}g^2, \cdot, \cdot)$	$(\mathbb{Z}/(13), \frac{2}{13}g^2)$	$1-\overline{\zeta_3}$	3	$\frac{1+\sqrt{13}}{2}$
$\mathcal{NG}((\mathbb{Z}/(3))^2, \frac{1}{3}(g^2-h^2), \cdot, \cdot)$	$(\mathbb{Z}/(13), \frac{2}{13}g^2)$	3	3	$\frac{1+\sqrt{13}}{2}$

$\lvert G\rvert = 11$	(G', q')	$\nu_3(\rho)$	$\nu_5(\rho)$	$\nu_{11}(\rho)$	$\nu_{15}(\rho)$
$\mathcal{NG}(\mathbb{Z}/(11), \frac{1}{11}g^2, \cdot, \zeta_{12}^7)$	$(\mathbb{Z}/(15), \frac{2}{15}g^2)$	$\frac{1-i\sqrt{3}}{2}$	$\frac{1+\sqrt{5}}{2}$	$\frac{11-i\sqrt{11}}{2}$	$\frac{1+i\sqrt{15}}{2}$
$\mathcal{NG}(\mathbb{Z}/(11), \frac{1}{11}g^2, \cdot, \overline{\zeta_{12}})$	$(\mathbb{Z}/(15), \frac{1}{15}g^2)$	$\frac{1+i\sqrt{3}}{2}$	$\frac{1-\sqrt{5}}{2}$	$\frac{11-i\sqrt{11}}{2}$	$\frac{1+i\sqrt{15}}{2}$
$\mathcal{NG}(\mathbb{Z}/(11), -\frac{1}{11}g^2, \cdot, \zeta_{12})$	$(\mathbb{Z}/(15), -\frac{1}{15}g^2)$	$\frac{1-i\sqrt{3}}{2}$	$\frac{1-\sqrt{5}}{2}$	$\frac{11+i\sqrt{11}}{2}$	$\frac{1-i\sqrt{15}}{2}$
$\mathcal{NG}(\mathbb{Z}/(11), -\frac{1}{11}g^2, \cdot, \zeta_{12}^5)$	$(\mathbb{Z}/(15), -\frac{2}{15}g^2)$	$\frac{1+i\sqrt{3}}{2}$	$\frac{1+\sqrt{5}}{2}$	$\frac{11+i\sqrt{11}}{2}$	$\frac{1-i\sqrt{15}}{2}$

$\lvert G\rvert = 13$	(G', q')	$\nu_{13}(\rho)$	$\nu_{17}(\rho)$
$\mathcal{NG}(\mathbb{Z}/(13), \frac{1}{13}g^2, b_1, -1)$	$(\mathbb{Z}/(17), \frac{3}{17}g^2)$	$\frac{13-\sqrt{13}}{2}$	$\frac{1+\sqrt{17}}{2}$
$\mathcal{NG}(\mathbb{Z}/(13), \frac{1}{13}g^2, b_2, -1)$	$(\mathbb{Z}/(17), \frac{3}{17}g^2)$	$\frac{13-\sqrt{13}}{2}$	$\frac{1+\sqrt{17}}{2}$
$\mathcal{NG}(\mathbb{Z}/(13), \frac{2}{13}g^2, b_3, 1)$	$(\mathbb{Z}/(17), \frac{1}{17}g^2)$	$\frac{13+\sqrt{13}}{2}$	$\frac{1-\sqrt{17}}{2}$
$\mathcal{NG}(\mathbb{Z}/(13), \frac{2}{13}g^2, b_4, 1)$	$(\mathbb{Z}/(17), \frac{1}{17}g^2)$	$\frac{13+\sqrt{13}}{2}$	$\frac{1-\sqrt{17}}{2}$

Remark 4.12. See that for $G = \mathbb{Z}/(13)$ we have two pairs of inequivalent fusion categories with the same indicators; hence, the near-group fusion ring $\mathrm{NG}(\mathbb{Z}/(13), 13)$ does not have FS indicator rigidity. Note that the lesser odd order groups *do* exhibit FS indicator rigidity.

5. Frobenius–Schur indicators for Haagerup–Izumi fusion categories

Near groups are examples of *quadratic* fusion categories: those tensor-generated by a single noninvertible simple object ρ where the set of simple objects is given by

$$G \cup \{\bar{g} \otimes \rho \mid \bar{g} \text{ a coset representative in } G/H\}$$

where H is some subgroup of G. Near groups correspond to $H = G$. On the other end of the spectrum, the Haagerup–Izumi fusion categories correspond to $H = \{e\}$.

Definition 5.1. $\mathcal{C}$ is a *Haagerup–Izumi* fusion category if

$$K_0(\mathcal{C}) = \mathrm{HI}(G) := \mathbb{Z}[G \cup \{g\rho \mid g \in G\}]$$

where multiplication is given by the group law and

$$g(h\rho) = (gh)\rho = (h\rho)g^{-1},$$
$$(g\rho)(h\rho) = gh^{-1} + \sum_{x \in G} x\rho.$$

Classification and examples. The complete lists of Haagerup–Izumi categories for $G = \mathbb{Z}/(3)$ and $G = \mathbb{Z}/(5)$ were found by Evans and Gannon [2017] without assuming unitarity by generalizing Izumi's methods to endomorphisms of Leavitt algebras. These categories are classified up to isomorphism of the group G and the parameters

- a sign $\pm$,
- ω a third root of unity, and
- $A \in M_{|G|}(\mathbb{C})$ a complex matrix

all satisfying some relations given in [Evans and Gannon 2017, Theorem 1]. An Haagerup–Izumi category with the above parameters will be denoted

$$\mathcal{HI}(G, \pm, \omega, A).$$

The notion of equivalence for the parameters is given in [Evans and Gannon 2017, Theorem 2(b)]. In particular, the category is unitary if and only if both the sign is $+$ and A is hermitian [Evans and Gannon 2017, Theorem 2(c)].

The most important examples of HI categories are the Yang–Lee system of sectors, which is the unique nonunitary such category with G the trivial group, and the system of sectors for the Haagerup subfactor, which is a unitary HI category with $G = \mathbb{Z}/(3)$.

Indicators when |G| is odd. When $\mathcal{C}$ is a Haagerup–Izumi fusion category the modular data for the center $\mathcal{Z}(\mathcal{C})$ was computed by Evans and Gannon [2017, §6.3] and is given as follows:

$X \in \mathcal{I}rr(\mathcal{Z}(\mathcal{C}))$		$F(X)$	braiding	θ_X
$\mathbb{1}$		$\mathbb{1}$	1	1
B		$\mathbb{1} + \sum_{g\in G} g \otimes \rho$	1	1
$A_\psi = A_{\overline{\psi}}$		$2\mathbb{1} + \sum_{g\in G} g \otimes \rho$	$\psi \in \widehat{G} \setminus \{1\}$	1
$C_\phi^{(h)}$	$(h \in G_+)$	$h + h^{-1} + \sum_{g\in G} g \otimes \rho$	$\phi \in \widehat{G}$	$\phi(h)$
D_j	$(1 \le j \le \frac{1}{2}(\lvert G\rvert^2 + 3))$	$\sum_{g\in G} g \otimes \rho$		ζ_j

In the preceding table G_+ is defined by a partition $G = G_+ \sqcup \{e\} \sqcup G_-$ where $(G_+)^{-1} = G_-$, which is always possible since $|G|$ is odd.

The ζ_j are a solutions to a system of equations with coefficients given by $\pm$, ω, A. These equations are (6.14) and (6.16)–(6.19) in [Evans and Gannon 2017, §6.2]. See [Evans and Gannon 2017, Proposition 2]. For G odd order they make another conjecture:

Conjecture 5.2 [Evans and Gannon 2017, Conjecture 1]. *Suppose $|G|$ is odd. Then there exists a metric group (H, q'') of order $|G|^2 + 4 = 2m + 1$ such that the simple objects D_j are indexed by $h \in H \setminus \{e\}$ where $D_h = D_{h^{-1}}$ and*

$$\theta_{D_h} = e^{2\pi i m q''(h)}.$$

Remark 5.3. Evans and Gannon [2017, Theorem 3] show that the conjecture is true for Haagerup–Izumi fusion categories with $|G| = 1, 3, 5$.

Theorem 5.4. *Suppose $\mathcal{C}$ is a Haagerup–Izumi fusion category with $|G|$ odd satisfying Conjecture 5.2. Then*

$$\nu_k(\rho) = \tfrac{1}{2}\theta_k^G(e) + \tfrac{1}{2}\Theta(H, kmq'').$$

Proof. Let $d = \operatorname{qdim}(\rho)$ in the category $\mathcal{C}$. Again by using Theorem 2.5

$$\nu_k(\rho) = \frac{1}{\operatorname{qdim}(\mathcal{C})}\Bigg(\operatorname{qdim}(B) + \sum_{\overline{\psi}\neq\psi\in\widehat{G}} \operatorname{qdim}(A_\psi) + \sum_{e\neq h^{-1}\neq h\in G}\sum_{\phi\in\widehat{G}} \theta_{C_\phi^h}^k \operatorname{qdim}(C_\phi^h) + \sum_{\gamma^{-1}\neq\gamma\in H} \theta_{D_\gamma} \operatorname{qdim}(D_\gamma)\Bigg).$$

Letting $|G| = 2n + 1$ and $|H| = 2m + 1$ we may enumerate these odd order groups as

$$G = \{e, g_i, g_i^{-1} \mid 1 \le i \le n\} \quad \text{and} \quad H = \{e, h_j, h_j^{-1} \mid 1 \le j \le m\},$$

which gives us

$$\nu_k(\rho) = \frac{1}{\operatorname{qdim}(\mathcal{C})}\Bigg(|G| + |G|d + |G|dn + (2 + |G|d)\sum_{g_i,\phi} \phi(g_i)^k + |G|d \sum_{h_j\in H} \zeta_j^k\Bigg).$$

By the same argument as in the proofs of Theorems 3.2 and 4.8 we can see

$$\sum_{g_i,\phi} \phi(g_i)^k = \frac{|G|}{2}(\theta_k^G(e) - 1).$$

Hence, using the expression for the center's ribbon structure from Conjecture 5.2 and the fact that $|H| = |G|^2 + 4$ and $\operatorname{qdim}(\mathcal{C}) = 2|G| + d|G|^2$ we see

$$\begin{aligned}
\nu_k(\rho) &= \frac{|G|}{\operatorname{qdim}(\mathcal{C})}\left(\frac{2+|G|d}{2}\theta_k^G(e) + d + dn - \frac{|G|d}{2} + d\sum_{h_j\in H} e^{2\pi i k m q''(h_j)}\right) \\
&= \tfrac{1}{2}\theta_k^G(e) + \frac{|G|}{2\operatorname{qdim}(\mathcal{C})}\big(2d + 2dn - |G|d + d(\sqrt{|H|}\Theta(H, kmq'') - 1)\big) \\
&= \tfrac{1}{2}\theta_k^G(e) + \frac{|G|d\sqrt{|G|^2+4}}{2\operatorname{qdim}(\mathcal{C})}\Theta(H, kmq'') \\
&= \tfrac{1}{2}\theta_k^G(e) + \tfrac{1}{2}\Theta(H, kmq'').
\end{aligned}$$

□

Now we collect in the tables below the values of the Frobenius–Schur indicators for the Haagerup–Izumi categories constructed in [Evans and Gannon 2011]:

$G = \mathbb{Z}/(3)$	(H, q'')	$\nu_k(\rho)$	$\nu_3(\rho)$	$\nu_{13}(\rho)$
$\mathcal{HI}(\mathbb{Z}/(3), +, 1, A_1)$	$(\mathbb{Z}/(13), \frac{1}{13}g^2)$	$\frac{1}{2}(1 - (\frac{1}{13}k))$	1	$\frac{1+\sqrt{13}}{2}$
$\mathcal{HI}(\mathbb{Z}/(3), +, 1, A_2)$	$(\mathbb{Z}/(13), \frac{1}{13}g^2)$	$\frac{1}{2}(1 - (\frac{1}{13}k))$	1	$\frac{1+\sqrt{13}}{2}$
$\mathcal{HI}(\mathbb{Z}/(3), -, 1, A_3)$	$(\mathbb{Z}/(13), \frac{2}{13}g^2)$	$\frac{1}{2}(1 + (\frac{1}{13}k))$	2	$\frac{1+\sqrt{13}}{2}$
$\mathcal{HI}(\mathbb{Z}/(3), -, 1, A_4)$	$(\mathbb{Z}/(13), \frac{2}{13}g^2)$	$\frac{1}{2}(1 + (\frac{1}{13}k))$	2	$\frac{1+\sqrt{13}}{2}$

In the preceding table the integer k must be relatively prime to $3 \cdot 13 = 39$.

$G = \mathbb{Z}/(5)$	(H, q'')	$\nu_k(\rho)$	$\nu_5(\rho)$	$\nu_{29}(\rho)$
$\mathcal{HI}(\mathbb{Z}/(5), +, 1, A_6)$	$(\mathbb{Z}/(29), \frac{1}{29}g^2)$	$\frac{1}{2}(1 - (\frac{1}{29}k))$	2	$\frac{1+\sqrt{29}}{2}$
$\mathcal{HI}(\mathbb{Z}/(5), +, 1, A_7)$	$(\mathbb{Z}/(29), \frac{1}{29}g^2)$	$\frac{1}{2}(1 - (\frac{1}{29}k))$	2	$\frac{1+\sqrt{29}}{2}$
$\mathcal{HI}(\mathbb{Z}/(5), -, 1, A_8)$	$(\mathbb{Z}/(29), \frac{2}{29}g^2)$	$\frac{1}{2}(1 + (\frac{1}{29}k))$	3	$\frac{1+\sqrt{29}}{2}$
$\mathcal{HI}(\mathbb{Z}/(5), -, 1, A_9)$	$(\mathbb{Z}/(29), \frac{2}{29}g^2)$	$\frac{1}{2}(1 + (\frac{1}{29}k))$	3	$\frac{1+\sqrt{29}}{2}$

In the preceding table the integer k must be relatively prime to $5 \cdot 13 = 65$.

Remark 5.5. See that for each of $\mathbb{Z}/(3)$ and $\mathbb{Z}/(5)$ we have two pairs of inequivalent fusion categories with the same indicators; hence, the Haagerup–Izumi fusion rings do not have FS rigidity.

Note that in this case as well as the $m = |G| = 13$ near-group case the pairs have centers with the same modular data (although it is not established whether the centers are equivalent). In view of this we formulate the following conjecture.

Conjecture 5.6. *Two fusion categories with a given Grothendieck ring that are also Morita equivalent cannot be distinguished by their Frobenius–Schur indicators.*

Acknowledgments

The author wishes to thank his PhD advisor Susan Montgomery; this article comprises the results of the author's dissertation. The author expresses much gratitude to Masaki Izumi, Richard Ng, and Peter Schauenburg for many useful conversations during the development of this work. Thanks are also due to David Penneys for pointing out the paper [Evans and Gannon 2014] to the author and to Vaughan Jones for introducing the author to the work of Izumi.

References

[Bakalov and Kirillov 2001] B. Bakalov and A. Kirillov, Jr., *Lectures on tensor categories and modular functors*, University Lecture Series **21**, Amer. Math. Soc., Providence, RI, 2001. MR Zbl

[Basak and Johnson 2015] T. Basak and R. Johnson, "Indicators of Tambara–Yamagami categories and Gauss sums", *Algebra Number Theory* **9**:8 (2015), 1793–1823. MR Zbl

[Eilenberg and Mac Lane 1953] S. Eilenberg and S. Mac Lane, "On the groups $H(\Pi, n)$, I", *Ann. of Math.* (2) **58** (1953), 55–106. MR Zbl

[Eilenberg and Mac Lane 1954] S. Eilenberg and S. Mac Lane, "On the groups $H(\Pi, n)$, II: Methods of computation", *Ann. of Math.* (2) **60** (1954), 49–139. MR Zbl

[Etingof et al. 2004] P. Etingof, S. Gelaki, and V. Ostrik, "Classification of fusion categories of dimension pq", *Int. Math. Res. Not.* **2004**:57 (2004), 3041–3056. MR Zbl

[Etingof et al. 2015] P. Etingof, S. Gelaki, D. Nikshych, and V. Ostrik, *Tensor categories*, Mathematical Surveys and Monographs **205**, Amer. Math. Soc., Providence, RI, 2015. MR Zbl

[Evans and Gannon 2011] D. E. Evans and T. Gannon, "The exoticness and realisability of twisted Haagerup–Izumi modular data", *Comm. Math. Phys.* **307**:2 (2011), 463–512. MR Zbl

[Evans and Gannon 2014] D. E. Evans and T. Gannon, "Near-group fusion categories and their doubles", *Adv. Math.* **255** (2014), 586–640. MR Zbl

[Evans and Gannon 2017] D. E. Evans and T. Gannon, "Non-unitary fusion categories and their doubles via endomorphisms", *Adv. Math.* **310** (2017), 1–43. MR Zbl

[Isaacs 1976] I. M. Isaacs, *Character theory of finite groups*, Pure and Applied Mathematics **69**, Academic, New York, 1976. MR Zbl

[Izumi 2000] M. Izumi, "The structure of sectors associated with Longo–Rehren inclusions, I: General theory", *Comm. Math. Phys.* **213**:1 (2000), 127–179. MR Zbl

[Izumi 2001] M. Izumi, "The structure of sectors associated with Longo–Rehren inclusions, II: Examples", *Rev. Math. Phys.* **13**:5 (2001), 603–674. MR Zbl

[Izumi 2017] M. Izumi, "A Cuntz algebra approach to the classification of near-group categories", pp. 222–343 in *Proceedings of the 2014 Maui and 2015 Qinhuangdao conferences in honour of Vaughan F. R. Jones' 60th birthday*, edited by S. Morrison and D. Penneys, Proc. Centre Math. Appl. Austral. Nat. Univ. **46**, Austral. Nat. Univ., Canberra, 2017. MR Zbl

[Linchenko and Montgomery 2000] V. Linchenko and S. Montgomery, "A Frobenius–Schur theorem for Hopf algebras", *Algebr. Represent. Theory* **3**:4 (2000), 347–355. MR Zbl

[Mason and Ng 2005] G. Mason and S.-H. Ng, "Central invariants and Frobenius–Schur indicators for semisimple quasi-Hopf algebras", *Adv. Math.* **190**:1 (2005), 161–195. MR Zbl

[Müger 2003] M. Müger, "From subfactors to categories and topology, II: The quantum double of tensor categories and subfactors", *J. Pure Appl. Algebra* **180**:1–2 (2003), 159–219. MR Zbl

[Ng and Schauenburg 2007a] S.-H. Ng and P. Schauenburg, "Frobenius–Schur indicators and exponents of spherical categories", *Adv. Math.* **211**:1 (2007), 34–71. MR Zbl

[Ng and Schauenburg 2007b] S.-H. Ng and P. Schauenburg, "Higher Frobenius–Schur indicators for pivotal categories", pp. 63–90 in *Hopf algebras and generalizations*, edited by L. H. Kauffman et al., Contemp. Math. **441**, Amer. Math. Soc., Providence, RI, 2007. MR Zbl

[Ng and Schauenburg 2008] S.-H. Ng and P. Schauenburg, "Central invariants and higher indicators for semisimple quasi-Hopf algebras", *Trans. Amer. Math. Soc.* **360**:4 (2008), 1839–1860. MR Zbl

[Serre 1977] J.-P. Serre, *Linear representations of finite groups*, Graduate Texts Math. **42**, Springer, 1977. MR Zbl

[Shimizu 2011] K. Shimizu, "Frobenius–Schur indicators in Tambara–Yamagami categories", *J. Algebra* **332** (2011), 543–564. MR Zbl

[Suzuki and Wakui 2002] K. Suzuki and M. Wakui, "On the Turaev–Viro–Ocneanu invariant of 3-manifolds derived from the E_6-subfactor", *Kyushu J. Math.* **56**:1 (2002), 59–81. MR Zbl

[Tambara and Yamagami 1998] D. Tambara and S. Yamagami, "Tensor categories with fusion rules of self-duality for finite abelian groups", *J. Algebra* **209**:2 (1998), 692–707. MR Zbl

[Wall 1963] C. T. C. Wall, "Quadratic forms on finite groups, and related topics", *Topology* **2** (1963), 281–298. MR Zbl

Received June 9, 2017. Revised January 20, 2019.

HENRY TUCKER
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CA
UNITED STATES
hjtucker@ucsd.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 303, No. 1, 2019
dx.doi.org/10.2140/pjm.2019.303.361

COMPACTNESS THEOREMS FOR 4-DIMENSIONAL GRADIENT RICCI SOLITONS

YONGJIA ZHANG

We prove compactness theorems for noncompact 4-dimensional shrinking and steady gradient Ricci solitons, respectively, satisfying: (1) every bounded open subset can be embedded in a closed 4-manifold with vanishing second homology group, and (2) are strongly κ-noncollapsed on all scales with respect to a uniform κ. These solitons are of interest because they are the only ones that can arise as finite-time singularity models for a Ricci flow on a closed 4-manifold with vanishing second homology group.

1. Introduction

Since the works of Cheeger and Gromov, compactness and precompactness theorems have played a fundamental role in understanding the geometry and topology of Riemannian manifolds. In the setting of the Ricci flow, Shi's local derivative of curvature estimates [1989] enabled Hamilton [1995a] to improve the convergence to C^∞-convergence of solutions. In dimension 3, in the setting of ancient noncollapsed Ricci flow, this was remarkably strengthened by Perelman [2002] who proved that the global curvature and bound follows from a curvature bound only at a single point. In dimensions 4 and above, this is no longer possible because of the existence of asymptotically conical singularity models, and in particular, asymptotically conical shrinking gradient Ricci solitons. Besides the weakness of the hypotheses, one of the strengths of Perelman's compactness theorem is that it is indeed a compactness result, not just a precompactness result. So the limit extracted from a subsequence is in the same class of objects as the original sequence of objects, in Perelman's case, 3-dimensional ancient κ-solutions.

In this paper we consider 4-dimensional Ricci solitons satisfying a certain topological condition which is of interest in the study of the Ricci flow on closed 4-manifolds with vanishing second homology group, which include homotopy 4-spheres. In singularity analysis of the Ricci flow in relation to developing a theory of Ricci flow with surgery, one considers the case where the underlying manifold of the Ricci shrinker is noncompact. In view of this, we seek a *pointed* compactness result.

MSC2010: 53C44.
Keywords: dimension four, Ricci soliton, compactness.

The large body of works by Munteanu and Wang on gradient Ricci solitons [2011; 2012; 2014; 2015; 2016; 2017a] led them to conjecture that 4-dimensional Ricci shrinkers may be classifiable. Indeed, this classification is completed under the condition of nonnegative isotropic curvature, or nonnegative sectional curvature, or nonnegative curvature operator; see [Li et al. 2018; Munteanu and Wang 2017b; Naber 2010]. In the more general case, Munteanu and Wang have made substantial progress towards their conjecture that such objects either are the quotients of splitting Ricci shrinkers or are asymptotically conical Ricci shrinkers. In the most optimistic version of their conjecture, one would expect that a *generic* Ricci flow with surgery on a closed 4-manifold would only produce a quotient 2-surgery, a quotient 3-surgery, or a smooth blow down, all in the case of a type I singularity. More conservatively, one may not wish to rule out Ricci flat ALE spaces and cohomegeneity-one steady gradient Ricci solitons forming generically as singularity models in dimension 4. Returning to dimension 3, paradoxically Perelman's theory of the space of noncompact ancient κ-solutions with positive sectional curvature, which builds on the work of Hamilton [1995b] and which is one of the deepest in the subject, is about a space conjectured by Perelman to be only a single point, namely the Bryant soliton. Brendle's proof of the uniqueness of the Bryant soliton in the class of nonflat 3-dimensional κ-noncollapsed steady Ricci solitons is also a deep result; see [Brendle 2013]. For these reasons, one may expect that a 4-dimensional theory of singularity models for Ricci flow may be related to the prototypical cases (perhaps more so than in dimension 3), which are the shrinking and steady Ricci solitons.

A triple (M^n, g, f), where (M^n, g) is a Riemannian manifold and f is a function on M^n, is called a gradient Ricci soliton, if

$$\mathrm{Ric} + \nabla^2 f = \frac{\lambda}{2} g,$$

where λ is a constant and when $\lambda > 0$, $\lambda = 0$ or $\lambda < 0$ the soliton is called shrinking, steady or expanding, respectively. In this paper we focus on shrinking and steady gradient Ricci solitons, or Ricci shrinkers and Ricci steadies for short, respectively. In other words, we always let $\lambda \geq 0$. By scaling the metric and adding a constant to the potential function f, a Ricci shrinker can always be normalized in the following way:

$$\begin{aligned} \mathrm{Ric} + \nabla^2 f &= \tfrac{1}{2} g, \\ |\nabla f|^2 + R &= f, \end{aligned} \tag{1-1}$$

and a non-Ricci-flat Ricci steady can be normalized in the following way:

$$\begin{aligned} \mathrm{Ric} + \nabla^2 f &= 0, \\ |\nabla f|^2 + R &= 1. \end{aligned} \tag{1-2}$$

Shrinking and steady gradient Ricci solitons are of great interest in the study of the singularity formation for the Ricci flow. For instance, they arise as blow-up limits

of finite-time singularities in Ricci flows; see [Enders et al. 2011; Gu and Zhu 2008; Hamilton 1995b], and Ricci shrinkers are also blow-down limits of ancient solutions with nonnegative curvature operator [Perelman 2002]. In this paper, we restrict our attention to the shrinking and steady gradient Ricci solitons satisfying a topological assumption, that is, every bounded open subset can be embedded in a closed 4-manifold with vanishing second homology group. This condition was previously considered by Bamler and Zhang [2017]. Besides that, we impose a uniform strong noncollapsing assumption, which fortunately holds for singularity models; see below. We define the following space of Ricci shrinkers and Ricci steadies.

Definition 1.1. Given $\kappa > 0$, $\mathcal{M}^4(\kappa)$ is the collection of all the 4-dimensional noncompact shrinking gradient Ricci solitons (M^4, g, f, p), where p is the point at which f attains its minimum, satisfying:

(a) (M^4, g) is nonflat.

(b) Every bounded open subset of M^4 can be embedded in a closed 4-manifold N^4 with $H_2(N) = 0$, where H_2 is the second homology group with coefficients in $\mathbb{Z}$.

(c) (M^4, g) is strongly κ-noncollapsed on all scales.

Definition 1.2. Given $\kappa > 0$, $\mathcal{N}^4(\kappa)$ is the collection of all the 4-dimensional noncompact steady gradient Ricci solitons (M^4, g, f, p), where $p \in M$ is such that $f(p) = 0$, satisfying:

(a) (M^4, g) is nonflat.

(b) Every bounded open subset of M^4 can be embedded in a closed 4-manifold N^4 with $H_2(N) = 0$, where H_2 is the second homology group with coefficients in $\mathbb{Z}$.

(c) (M^4, g) is strongly κ-noncollapsed on all scales.

Remarks: (1) In item (b) of Definition 1.1 and 1.2, one may simply assume that M^4 can be embedded in N^4 and the same curvature estimates in Section 4 still hold. However, our assumption is more natural in view of singularity models; see below for more details.

(2) The closed 4-manifold N^4 mentioned in item (b) of both Definition 1.1 and 1.2 may depend on the soliton (M^4, g, f, p) or even the open subset, we do not need to assume that every soliton in $\mathcal{M}^4(\kappa)$ or $\mathcal{N}^4(\kappa)$ satisfies this property for the same N^4.

(3) In Definition 1.1 the base point p is the always the minimum point of the potential function f, whereas in Definition 1.2 the base point p can be fixed at any point in M, since one can always replace f by $f - f(p)$, and this does not affect the normalization (1-2)

(4) Since Ricci-flatness and strong noncollapsing condition implies ALE (Corollary 8.86 in [Cheeger and Naber 2015]), by Theorem 6.1 such ALE manifold, when regarded as Ricci steadies, cannot be included in $\mathcal{N}^4(\kappa)$. Henceforth, unless otherwise stated, we always work on non-Ricci-flat Ricci steadies.

(5) There are only a few examples for simply connected 4-dimensional Ricci shrinkers: $\mathbb{S}^4$, $\mathbb{S}^2 \times \mathbb{R}^2$, $\mathbb{S}^3 \times \mathbb{R}$, $\mathbb{S}^2 \times \mathbb{S}^2$, and the FIK shrinker (see [Feldman et al. 2003]). Noncollapsed simply connected 4-dimensional Ricci steady has more examples, except for the Bryant soliton [2005], there is a family of Ricci steadies discovered by Appleton [2017]. However, Appleton's solitons are not κ-noncollapsed with respect to a uniform κ.

By strong noncollapsing we mean the following:

Definition 1.3. A Riemannian manifold (M^n, g) is strongly κ-noncollapsed on all scales, where $\kappa > 0$, if the following holds. For all $x \in M$ and $r > 0$, if $R < r^{-2}$ on $B(x, r)$, then $\mathrm{Vol}(B(x, r)) \geq \kappa r^n$. Here we use R to denote the scalar curvature.

Our main theorems are the following:

Theorem 1.4. *$\mathcal{M}^4(\kappa)$ is compact in the smooth pointed Cheeger–Gromov sense, where each $(M^4, g, f, p) \in \mathcal{M}^4(\kappa)$ is normalized as in* (1-1).

Theorem 1.5. *$\mathcal{N}^4(\kappa)$ is precompact in the smooth pointed Cheeger–Gromov sense, where each $(M^4, g, f, p) \in \mathcal{N}^4(\kappa)$ is normalized as in* (1-2). *Furthermore, for any convergent sequence in $\mathcal{N}^4(\kappa)$, the limit is either the Euclidean space or still lies in $\mathcal{N}^4(\kappa)$.*

Here by saying that $\mathcal{N}^4(\kappa)$ is precompact we mean that for every sequence $\{(M_k^4, g_k, f_k, p_k)\}_{k=1}^{\infty}$ contained in $\mathcal{N}^4(\kappa)$, there exists a subsequence that converges in the pointed smooth Cheeger–Gromov sense to a Ricci steady $(M_\infty^4, g_\infty, f_\infty, p_\infty)$; by saying that $\mathcal{M}^4(\kappa)$ is compact we mean that first of all it is precompact, and furthermore, the limit of every convergent sequence in $\mathcal{M}^4(\kappa)$ also lies in $\mathcal{M}^4(\kappa)$.

A homotopy four-sphere, as a particular example, has vanishing second homology group. When approaching the 4-dimensional smooth Poincaré conjecture using the Ricci flow with surgery, the Ricci solitons that may arise in the analysis of the first singularities, being the blow-up Cheeger–Gromov–Hamilton limit of the homotopy four-sphere, satisfies the property that every open bounded subset can be embedded in the original homotopy four-sphere. Furthermore, according to Perelman [2002], every Ricci flow on closed manifold forming a finite-time singularity is strongly κ-noncollapsed on some fixed finite positive scale, where $\kappa > 0$ depends only on the initial data, the length of the time interval of the Ricci flow, and the scale (see Theorem 6.74 in [Chow et al. 2007]). Thus any blow-up limit at the singular time must be strongly κ-noncollapsed on all scales. Therefore, all Ricci shrinkers or Ricci steadies that arise from such singularity analysis must lie in $\mathcal{M}^4(\kappa)$ or $\mathcal{N}^4(\kappa)$,

respectively. We hope that our result will be helpful to the finite-time singularity analysis for the Ricci flow on 4-dimensional closed manifolds. We mention here that Hamilton [1995b] classified finite-time singularities as type I and type II, while Ricci shrinkers and Ricci steadies, being singularity models, are correspondent to these two singularity types, respectively. It is known that the fixed-point blow-up limit of a type I singularity is always a nonflat Ricci shrinker [Enders et al. 2011; Naber 2010], but it remains open whether a similar result is true for type II singularities, that is, *is the blow-up limit of a type II singularity, obtained by some careful point picking, always a Ricci steady?* Hamilton answered this question positively under the assumption that the blow-up limit, obtained by some careful point-picking, has nonnegative curvature operator; see [Hamilton 1993; 1995b].

Condition (b) in both Definition 1.1 and 1.2 plays a very important role in ruling out the Ricci-flat limits. By the strong noncollapsing property, a Ricci-flat blow-up limit of a Ricci flow at a finite-time singularity must have Euclidean volume growth, and, according to Cheeger and Naber [2015], must be *asymptotically locally Euclidean* (ALE for short), which cannot be embedded in any closed 4-manifold with vanishing second homology group (see Corollary 5.8 in [Anderson 2010]; an alternative proof by Richard Bamler is provided in Section 6). This idea gives a uniform curvature growth estimate for every element in the space $\mathcal{M}^4(\kappa)$ and a uniform curvature bound for every element in the space $\mathcal{N}^4(\kappa)$; see Theorems 4.1 and 4.2 below, from which we obtain the compactness results. This argument is in the spirit of Perelman's bounded curvature at bounded distance result for κ-solutions with nonnegative curvature operator (see section 11 of [Perelman 2002]). Perelman also assumes a uniform κ, which is motivated by the reason that all these κ-solutions arise from the same Ricci flow that forms a finite-time singularity. However, there is always a universal κ for all the 3-dimensional κ-solutions that is not a shrinking space form because of the classification of 3-dimensional Ricci shrinkers.

In their papers, Haslhofer and Müller [2011; 2015] have proved a compactness theorem for 4-dimensional Ricci shrinkers, where they only assume a uniform lower bound of the entropy, but where the limit could possibly be an orbifold shrinker. In comparison, the strong noncollapsing assumption in our theorem is correspondent to their bounded entropy assumption (indeed, it is clear that a uniform lower bound of entropy implies κ-noncollapsing with respect to a universal κ; see [Carrillo and Ni 2009; Yokota 2012], yet we do not know how it is related to our strong noncollapsing assumption); in addition, we have a topological restriction. What is novel in our work is that the orbifold Ricci shrinkers will never show up as limits.

From the proof of Theorem 1.4, we also get the following property of the space $\mathcal{M}^4(\kappa)$:

Corollary 1.6. *There exist $C_1 > 0$, $C_2 > 0$, and $C_3 < \infty$ depending only on κ, such that for every $(M^4, g, f, p) \in \mathcal{M}^4(\kappa)$ the following hold*:

(a) $R(p) > C_1$.

(b) $R(x) > C_2 f^{-1}(x)$, *for all* $x \in M$.

(c) $|\pi_1(M)| < C_3$, *where* $\pi_1(M)$ *is the fundamental group of* M.

This paper is organized as follows. In Section 2 we collect some known results for Ricci solitons, which are used in our arguments. In Section 3 we carry out some a priori estimates. In Section 4 we prove curvature estimates for the Ricci shrinkers in $\mathcal{M}^4(\kappa)$ and for the Ricci steadies in $\mathcal{N}^4(\kappa)$. In Section 5 we prove Theorems 1.4, 1.5, and Corollary 1.6. In Section 6 we provide an alternative proof of Anderson's theorem [2010].

2. Preliminaries

In this section we collect some well-known results that are used in our proof. Notice that in this paper the Ricci soliton equations that we work with may not be normalized as (1-1) or (1-2), since sometimes scaling is necessary. Hence we will specify the Ricci soliton equations in every statement. We start with the following differential equations for the geometric quantities on Ricci shrinkers and steadies.

Proposition 2.1. *Let* (M, g, f) *be a shrinking or steady gradient Ricci soliton satisfying*

$$\mathrm{Ric} + \nabla^2 f = \frac{\lambda}{2} g,$$

where $\lambda \geq 0$. *Then the following hold.*

$$\Delta_f R = \lambda R - 2|\mathrm{Ric}|^2, \tag{2-1}$$

$$\Delta_f \mathrm{Ric} = \lambda\, \mathrm{Ric} + \mathrm{Rm} * \mathrm{Ric}, \tag{2-2}$$

$$\Delta_f \mathrm{Rm} = \lambda\, \mathrm{Rm} + \mathrm{Rm} * \mathrm{Rm}, \tag{2-3}$$

$$\Delta_f \nabla^k \mathrm{Rm} = \lambda\Big(\frac{k}{2} + 1\Big)\nabla^k \mathrm{Rm} + \sum_{j=0}^{k} \nabla^j \mathrm{Rm} * \nabla^{k-j} \mathrm{Rm}, \tag{2-4}$$

where $*$ *stands for some contraction and* $\Delta_f = \Delta - \langle \nabla f, \nabla \cdot \rangle$ *is the* f*-Laplacian operator.*

Proof. Since every reference on these differential equations we can find deals only with the case $\lambda = 1$ or $\lambda = 0$, we take (2-3) as an example to quickly sketch how these formulae can be carried out; other equations can be proved in the same way. Recall that the canonical form of a Ricci soliton $g(t) = \tau(t)\varphi_t^*(g)$ evolves by the Ricci flow (see Theorem 4.1 of [Chow et al. 2006]), where

$$\tau(t) = 1 - \lambda t, \quad \frac{d}{dt}\varphi_t = \frac{1}{\tau}\nabla f \circ \varphi_t, \quad \varphi_0 = \mathrm{id}.$$

Taking Rm as a (4, 0)-tensor, we have $\mathrm{Rm}(g(t)) = \tau(t)\,\mathrm{Rm}(\varphi_t^*(g)) = \tau(t)\varphi_t^*(\mathrm{Rm})$. Hence by the standard curvature evolution equation (see Theorem 7.1 of [Hamilton 1982]) we have

$$\Delta\,\mathrm{Rm} + \mathrm{Rm} * \mathrm{Rm} = \frac{\partial}{\partial t}\Big|_{t=0} \mathrm{Rm}(g(t)) = -\lambda\,\mathrm{Rm} + \mathcal{L}_{\nabla f}\,\mathrm{Rm},$$

where $\mathcal{L}$ stands for the Lie derivative. Let Y_1, Y_2, Y_3, Y_4 be four arbitrary vector fields. Then

$$\begin{aligned}
\mathcal{L}_{\nabla f}\,\mathrm{Rm}(Y_1, Y_2, Y_3, Y_4) &= \nabla_{\nabla f}(\mathrm{Rm}(Y_1, Y_2, Y_3, Y_4)) - \mathrm{Rm}(\mathcal{L}_{\nabla f} Y_1, Y_2, Y_3, Y_4) \\
&\quad - \mathrm{Rm}(Y_1, \mathcal{L}_{\nabla f} Y_2, Y_3, Y_4) - \mathrm{Rm}(Y_1, Y_2, \mathcal{L}_{\nabla f} Y_3, Y_4) \\
&\quad - \mathrm{Rm}(Y_1, Y_2, Y_3, \mathcal{L}_{\nabla f} Y_4) \\
&= \nabla_{\nabla f}\,\mathrm{Rm}(Y_1, Y_2, Y_3, Y_4) + \mathrm{Rm}(\nabla_{Y_1}\nabla f, Y_2, Y_3, Y_4) \\
&\quad + \mathrm{Rm}(Y_1, \nabla_{Y_2}\nabla f, Y_3, Y_4) + \mathrm{Rm}(Y_1, Y_2, \nabla_{Y_3}\nabla f, Y_4) \\
&\quad + \mathrm{Rm}(Y_1, Y_2, Y_3, \nabla_{Y_4}\nabla f).
\end{aligned}$$

Taking into account that $\nabla^2 f = \frac{\lambda}{2} g - \mathrm{Ric}$ we obtain the conclusion. □

The following two propositions for the potential function growth rate and the volume growth rate for Ricci shrinkers were proved by Cao and Zhou [2010], with an observation of Munteanu [2009]. We use its sharpened version of Haslhofer and Müller [2011]. Besides that, Munteanu and Wang [2014] proved a volume growth estimate with an improved constant.

Proposition 2.2. *Let (M^n, g, f) be a noncompact shrinking gradient Ricci soliton normalized as in* (1-1). *Let p be a point where f attains its minimum. Then the following holds*:

$$\tfrac{1}{4}(d(x, p) - 5n)_+^2 \le f(x) \le \tfrac{1}{4}(d(x, p) + \sqrt{2n})^2, \tag{2-5}$$

where $u_+ := \max\{u, 0\}$ denotes the positive part of a function.

Proposition 2.3. *There exists $C < \infty$ depending only on the dimension n, such that under the same assumption of Proposition 2.2 the following holds*:

$$\mathrm{Vol}(B(p, r)) \le Cr^n, \tag{2-6}$$

for all $r > 0$.

To locally estimate the Ricci curvature, we need the following local Sobolev inequality, whose constant depends only on the local geometry.

Proposition 2.4. *For all $\kappa > 0$, there exists $C < \infty$ and $\delta \in (0, 2)$, depending only on κ and the dimension $n \ge 3$ such that the following holds. Let (M^n, g) be*

a Riemannian manifold and $x_0 \in M$, and assume that $|\mathrm{Rm}| \le 2$ on $B(x_0, 2)$ and $\mathrm{Vol}(B(x_0, 2)) \ge \kappa$. Then

$$\|u\|_{L^{2n/(n-2)}} \le C\|\nabla u\|_{L^2}, \tag{2-7}$$

for all $u \in C_0^\infty(B(x_0, \delta))$.

Proof. This is a standard result; for the convenience of the readers we sketch the proof. We follow the lines of reasoning of Lemma 3.2 of [Haslhofer and Müller 2011]. We need only to prove an L^1 Sobolev inequality

$$\|u\|_{L^{n/(n-1)}} \le C_1\|\nabla u\|_{L^1}, \tag{2-8}$$

for all $u \in C_0^\infty(B(x_0, \delta))$, where δ and C_1 depend only on κ and the dimension n. Then (2-7) follows from (2-8). Indeed, C_1 is equal to the isoperimetric constant of $B(x_0, \delta)$,

$$C_1 = C_I = \sup |\Omega|^{n/(n-1)}/|\partial\Omega|,$$

where the supremum is taken over all the open sets $\Omega \subset B(x_0, \delta)$ with smooth boundary. By a theorem of Croke [1980, Theorem 11], C_I can be estimated by

$$C_I \le C(n)\omega^{-(n+1)/n},$$

where $C(n)$ is a constant depending only on the dimension and ω is the visibility angle defined by

$$\omega = \inf_{y \in B(x_0,\delta)} |U_y|/|\mathbb{S}^{n-1}|,$$

where $U_y = \{v \in T_y B(x_0, \delta) : |v| = 1$, the geodesic γ_v minimizes up to $\partial B(x_0, \delta)\}$.

We restrict δ in $\left(0, \frac{1}{2}\right)$ and let y be an arbitrary point in $B(x_0, \delta)$. Let

$$J(r, \theta)dr \wedge d\theta, \quad \bar{J}(r, \theta)dr \wedge d\theta$$

be the volume elements in terms of spherical normal coordinates around the point y and in the hyperbolic space with constant sectional curvature -2, respectively. By the relative volume comparison theorem, we have

$$\begin{aligned} c_2\kappa - C_3\delta^n \le |B(x_0, 1)| - |B(x_0, \delta)| &\le \int_{U_y}\int_0^{1+\delta} J(r, \theta)\, dr\, d\theta \\ &\le \int_{U_y}\int_0^{1+\delta} \bar{J}(r, \theta)\, dr\, d\theta \le C_4|U_y|\left(\tfrac{3}{2}\right)^n, \end{aligned}$$

where c_2, C_3, and C_4 are constants depending only on the dimension n. Taking $\delta = (c_2\kappa/(2C_3))^{1/n}$, we have that $|U_y|$ is bounded from below by a constant depending only on κ and the dimension n, for all $y \in B(x_0, \delta)$, and the conclusion follows. □

We conclude this section with the following gap theorem of Yokota [2009; 2012], which is used in the proof of Theorem 1.4 to show that the limit shrinker is nonflat.

Proposition 2.5. *There exists $\varepsilon > 0$ depending only on the dimension n such that the following holds. Let (M^n, g, f) be a shrinking gradient Ricci soliton, which is normalized as in* (1-1). *If*

$$\frac{1}{(4\pi)^{n/2}} \int_M e^{-f}\, dg > 1 - \varepsilon,$$

then (M^n, g, f) is the Gaussian shrinker, that is, (M^n, g) is the Euclidian space.

3. A priori estimates

The a priori estimates in this section hold for any dimension $n \geq 3$. We start with a localized derivative estimate for the Riemann curvature tensor.

Proposition 3.1. *There exists $C < \infty$ depending only on the dimension n such that the following holds. Let (M^n, g, f) be a shrinking or steady gradient Ricci soliton such that*

$$\mathrm{Ric} + \nabla^2 f = \frac{\lambda}{2} g,$$

where $\lambda \geq 0$. Let $x_0 \in M$ and $r > 0$. If $|\mathrm{Rm}| \leq r^{-2}$ and $|\nabla f| \leq r^{-1}$ on $B(x_0, 2r)$, then

$$|\nabla \mathrm{Rm}| \leq C r^{-3} \text{ on } B(x_0, r). \tag{3-1}$$

More generally, there exist C_l depending only on $l \geq 0$ and the dimension n, such that under the above assumptions, it holds that

$$|\nabla^l \mathrm{Rm}| \leq C_l r^{-2-l} \text{ on } B(x_0, r). \tag{3-2}$$

Proof. The proof is a standard elliptic modification of Shi's estimates [1989]. One can also combine [Shi 1989] with the canonical form to obtain this result. For the readers' convenience we will give a proof for (3-1). The higher derivative estimates (3-2) follow in a standard way by induction. We compute using (2-3)

$$\begin{aligned}
\Delta_f |\mathrm{Rm}|^2 &= 2\langle \mathrm{Rm}, \Delta_f \mathrm{Rm}\rangle + 2|\nabla \mathrm{Rm}|^2 \\
&= 2|\nabla \mathrm{Rm}|^2 + 2\lambda |\mathrm{Rm}|^2 + \mathrm{Rm} * \mathrm{Rm} * \mathrm{Rm} \\
&\geq 2|\nabla \mathrm{Rm}|^2 - C_1 |\mathrm{Rm}|^3,
\end{aligned}$$

where $C_1 < \infty$ depends only on the dimension n. By (2-4), we have

$$\begin{aligned}
\Delta_f |\nabla \mathrm{Rm}|^2 &= 2\langle \nabla \mathrm{Rm}, \Delta_f \nabla \mathrm{Rm}\rangle + 2|\nabla^2 \mathrm{Rm}|^2 \\
&= 2|\nabla^2 \mathrm{Rm}|^2 + 3\lambda |\nabla \mathrm{Rm}|^2 + \mathrm{Rm} * \nabla \mathrm{Rm} * \nabla \mathrm{Rm} \\
&\geq 2|\nabla^2 \mathrm{Rm}|^2 - C_2 |\mathrm{Rm}||\nabla \mathrm{Rm}|^2,
\end{aligned}$$

where $C_2 < \infty$ depends only on the dimension n.

Defining $u = (\beta r^{-4} + |\mathrm{Rm}|^2)|\nabla \mathrm{Rm}|^2$, where $\beta > 0$ is a constant that we will specify later, we have

$$\begin{aligned}\Delta_f u &= |\nabla \mathrm{Rm}|^2 \Delta_f |\mathrm{Rm}|^2 + (\beta r^{-4} + |\mathrm{Rm}|^2)\Delta_f |\nabla \mathrm{Rm}|^2 + 2\langle \nabla |\mathrm{Rm}|^2, \nabla |\nabla \mathrm{Rm}|^2\rangle \\ &\geq 2|\nabla \mathrm{Rm}|^4 - C_1 |\nabla \mathrm{Rm}|^2 |\mathrm{Rm}|^3 \\ &\qquad + (\beta r^{-4} + |\mathrm{Rm}|^2)(2|\nabla^2 \mathrm{Rm}|^2 - C_2 |\mathrm{Rm}||\nabla \mathrm{Rm}|^2) \\ &\qquad - 8|\nabla \mathrm{Rm}| \cdot \big|\nabla |\mathrm{Rm}|\big| \cdot |\nabla^2 \mathrm{Rm}| \cdot |\mathrm{Rm}| \\ &\geq 2|\nabla \mathrm{Rm}|^4 - C_1 |\nabla \mathrm{Rm}|^2 |\mathrm{Rm}|^3 \\ &\qquad + (\beta r^{-4} + |\mathrm{Rm}|^2)(2|\nabla^2 \mathrm{Rm}|^2 - C_2 |\mathrm{Rm}||\nabla \mathrm{Rm}|^2) \\ &\qquad - \tfrac{1}{2}|\nabla \mathrm{Rm}|^4 - 32|\nabla^2 \mathrm{Rm}|^2 |\mathrm{Rm}|^2,\end{aligned}$$

where we have used Kato's inequality as well as the Cauchy–Schwarz inequality. Letting $\beta = 16$ and taking into account that $|\mathrm{Rm}| \leq r^{-2}$ in $B(x_0, 2r)$, we have

$$\Delta_f u \geq \tfrac{3}{2}|\nabla \mathrm{Rm}|^4 - C_3 r^{-6} |\nabla \mathrm{Rm}|^2 \geq |\nabla \mathrm{Rm}|^4 - C_4 r^{-12},$$

where we have used the Cauchy–Schwarz inequality, and C_3 and C_4 are constants depending only on n. By the definition of u we have $|\nabla \mathrm{Rm}|^4 \geq (r^8/289)u^2$; hence

$$\Delta_f u \geq c_5 r^8 u^2 - C_5 r^{-12}, \tag{3-3}$$

where c_5 and C_5 are constants depending only on n.

We let $\phi(x) = \varphi(d(x_0, x))$ be the cut-off function, where $\varphi(s) = 0$ for $s \geq 2r$, $\varphi(s) = 1$ for $s \in [0, r]$, and

$$0 \leq \varphi \leq 1, \quad -2r^{-1} \leq \varphi'(s) \leq 0, \quad |\varphi''(s)| \leq 2r^{-2} \tag{3-4}$$

for all $s \in [r, 2r]$. We compute

$$\begin{aligned}\Delta_f(u\phi^2) &= \phi^2 \Delta_f u + u\Delta_f \phi^2 + 2\langle \nabla u, \nabla \phi^2\rangle \\ &\geq c_5 r^8 u^2 \phi^2 - C_5 r^{-12}\phi^2 + 2\langle \nabla(u\phi^2), \nabla \log \phi^2\rangle - 8|\nabla \phi|^2 u + u\Delta_f \phi^2.\end{aligned} \tag{3-5}$$

The last two terms in (3-5) need to be estimated. We have

$$|\nabla \phi|^2 = \varphi'^2 |\nabla d|^2 \leq 4r^{-2},$$

and

$$\begin{aligned}\Delta_f \phi^2 &= 2\phi(\varphi' \Delta_f d + \varphi''|\nabla d|^2) + 2\varphi'^2 |\nabla d|^2 \\ &= 2\phi(\varphi' \Delta d - \varphi' \langle \nabla f, \nabla d\rangle) + 2\phi\varphi''|\nabla d|^2 + 2\varphi'^2 |\nabla d|^2 \\ &\geq 2\left(-\frac{2(n-1)\coth(1)}{r^2} - \frac{2}{r^2}\right) - \frac{4}{r^2} \geq -C_6 r^{-2},\end{aligned}$$

where C_6 is a positive constant depending only on the dimension n. In the above derivation we have used $|\nabla f| \leq r^{-1}$, the properties of φ (3-4), the Laplacian

comparison theorem, and that $\varphi'(s) = 0$ for all $s \in [0, r]$. Inserting the above inequalities into (3-5), and defining $G = u\phi^2$, we have

$$\Delta_f G \geq c_5 r^8 \frac{G^2}{\phi^2} - C_5 r^{-12}\phi^2 + 2\langle \nabla G, \nabla \log \phi^2 \rangle - C_7 r^{-2}\frac{G}{\phi^2}. \tag{3-6}$$

Let $x_1 \in B(x_0, 2r)$ be a point where G attains its maximum. Taking into account that $0 \leq \phi \leq 1$, it follows from (3-6) that

$$c_5 r^8 G(x_1)^2 - C_7 r^{-2} G(x_1) - C_5 r^{-12} \leq 0,$$

which solves $G(x_1) \leq C_8 r^{-10}$, where C_7 and C_8 depend only on n. Therefore $u(x) \leq C_8 r^{-10}$ on $B(x_0, r)$, where $\phi = 1$ and $G = u$. It follows from the definition of the function u that

$$|\nabla \mathrm{Rm}|^2 \leq C r^{-6} \text{ on } B(x_0, r). \qquad \square$$

The following proposition says that the smallness of the scalar curvature on a ball implies the smallness of the Ricci curvature on a smaller ball. Our argument is inspired by Theorem 3.2 in [Wang 2012]. The same idea was implemented in [Bamler and Zhang 2017].

Proposition 3.2. *For any $\kappa > 0$, there exists $\delta \in (0, 2)$ and $C < \infty$, depending only on κ and the dimension n, such that the following holds. Let (M^n, g, f) be a shrinking or steady gradient Ricci soliton such that*

$$\mathrm{Ric} + \nabla^2 f = \frac{\lambda}{2} g,$$

where $\lambda \geq 0$. Let $x_0 \in M$ and $r \in (0, 1]$. If

$$|\mathrm{Rm}| \leq 2, \quad R \leq r^2, \quad |\nabla f| \leq r \text{ on } B(x_0, 2) \quad \text{and} \quad \mathrm{Vol}(B(x_0, 2)) \geq \kappa,$$

then

$$|\mathrm{Ric}| \leq Cr \text{ on } B\Big(x_0, \frac{\delta}{2}\Big). \tag{3-7}$$

Proof. We define a cut-off function that is similar to the one that we have used in the proof of the last proposition. Let $\phi(x) = \varphi(d(x_0, x))$, where $\varphi(s) = 0$ for $s \geq 2$, $\varphi(s) = 1$ for $s \in [0, 1]$, and

$$0 \leq \varphi \leq 1, \quad -2 \leq \varphi'(s) \leq 0, \quad |\varphi''(s)| \leq 2, \tag{3-8}$$

for $s \in [1, 2]$. Integrating (2-1) against ϕ, we have

$$\begin{aligned} 2\int |\mathrm{Ric}|^2 \phi &= \lambda \int R\phi - \int \phi \Delta R + \int \langle \nabla f, \nabla R \rangle \phi \\ &= \lambda \int R\phi - \int R\Delta\phi - \int \phi R \Delta f - \int R\langle \nabla \phi, \nabla f \rangle \end{aligned}$$

$$= \Big(1 - \frac{n}{2}\Big)\lambda \int R\phi - \int R\Delta\phi + \int \phi R^2 - \int R\langle \nabla\phi, \nabla f\rangle$$
$$\leq - \int R\Delta\phi + \int \phi R^2 - \int R\langle \nabla\phi, \nabla f\rangle,$$

where we have used $\Delta f = \frac{n}{2}\lambda - R$ and Chen's result [2009] that $R \geq 0$ on a Ricci shrinker or Ricci steady. By the Laplacian comparison theorem, the Bishop–Gromov volume comparison theorem, and the property of ϕ (3-8), we have

$$-\Delta\phi \leq C_1, \quad |\langle \nabla\phi, \nabla f\rangle| \leq C_2 r, \quad \mathrm{Vol}(B(x_0, 2)) \leq C_3,$$

where C_1, C_2, and C_3 are positive constants depending only on the dimension n. It then follows that

$$\int |\mathrm{Ric}|^2\phi \leq C_1 C_3 r^2 + C_3 r^4 + C_2 C_3 r^3,$$

and that

$$\|\mathrm{Ric}\|_{L^2(B(x_0,\delta))} \leq C_4 r,$$

where C_4 depends only on the dimension n, and $\delta \in (0, 2)$ is the positive number given by Proposition 2.4 that depends only on κ and the dimension n.

We have the following inequality satisfied by $|\mathrm{Ric}|$:

$$\begin{aligned} 2|\mathrm{Ric}|\Delta_f|\mathrm{Ric}| + 2\big|\nabla|\mathrm{Ric}|\big|^2 = \Delta_f|\mathrm{Ric}|^2 &= 2\langle \mathrm{Ric}, \Delta_f\,\mathrm{Ric}\rangle + 2|\nabla\,\mathrm{Ric}|^2 \\ &= 2\lambda|\mathrm{Ric}|^2 + 2|\nabla\,\mathrm{Ric}|^2 - \mathrm{Rm} * \mathrm{Ric} * \mathrm{Ric} \\ &\geq 2|\nabla\,\mathrm{Ric}|^2 - \mathrm{Rm} * \mathrm{Ric} * \mathrm{Ric}\,. \end{aligned}$$

Taking into account that $|\mathrm{Rm}| \leq 2$ on $B(x_0, 2)$ and Kato's inequality that $\big|\nabla|\mathrm{Ric}|\big|^2 \leq |\nabla\,\mathrm{Ric}|^2$, we have

$$\Delta_f|\mathrm{Ric}| \geq -C_5|\mathrm{Ric}|, \tag{3-9}$$

where C_5 depends on the dimension n. We use the local Sobolev inequality (2-7) to apply the standard Moser iteration to the inequality (3-9). Notice that we need to use $|\nabla f| \leq r \leq 1$ when performing the iteration. Indeed, this is the only reason why we have to put a restriction on the scale r. It follows that

$$\sup_{B(x_0,\,\delta/2)} |\mathrm{Ric}| \leq C_6\|\mathrm{Ric}\|_{L^2(B(x,\delta))} \leq Cr,$$

where C depends only on κ and the dimension n. □

4. Curvature estimates

In this section we prove a bounded curvature at bounded distance theorem for Ricci shrinkers in the space $\mathcal{M}^4(\kappa)$ as well as a uniformly bounded curvature

theorem for Ricci steadies in $\mathcal{N}^4(\kappa)$. These results are analogues to Perelman's bounded curvature at bounded distance result (see section 11 of [Perelman 2002]). The fact that the Ricci-flat limit does not appear in our argument plays a role as equally important as the fact that the asymptotic volume ratio equals zero in Perelman's argument. However, our results are somewhat weaker than Perelman's. In Theorem 4.1 we are only able to fix the base point where the potential function attains its minimum (or wherever is at a bounded distance to it), while in Theorem 4.2 the curvature bound is at a fixed scale instead of a relative scale, that is, the curvature bound is in terms of a fixed number instead of the curvature at an arbitrary base point. The reason in analysis is the following: to implement results in Section 3 in an argument of contradiction, the curvature largeness should be characterized by $|\nabla f|^2$. Suppose around a point the curvature is large in some relative sense but small compared to $|\nabla f|^2$. Then the a priori estimates we have established in Section 3 do not hold any more, since the assumption $|\nabla f| \le r \le 1$ made in Proposition 3.2 is no longer valid after scaling, and Moser iteration does not yield a nice bound for the Ricci curvature as in (3-7). To give a geometric understanding for the aforementioned weakness, we take an asymptotic conical shrinker as an example: one could take a sequence of points tending to infinity in an asymptotically conical Ricci shrinker, and the associated pointed limit is the asymptotic cone of the Ricci shrinker. Since this asymptotic cone is singular at its vertex, we have neither Perelman's bounded curvature at bounded distance nor compactness.

Theorem 4.1. *There exists $C < \infty$ and $D < \infty$ depending only on κ, such that the following holds. Let $(M^4, g, f, p) \in \mathcal{M}^4(\kappa)$ be normalized as in* (1-1)*. Then*

$$|\mathrm{Rm}|(x) \le C \quad \text{if } x \in B(p, 200),$$
$$\frac{|\mathrm{Rm}|}{f}(x) \le D \quad \text{if } x \notin B(p, 200).$$

Proof. We argue by contradiction. Suppose the statement is not true, then there exist a sequence of counterexamples $\{(M_k^4, f_k, g_k, p_k)\}_{k=1}^{\infty} \subset \mathcal{M}^4(\kappa)$ normalized as in (1-1), and $x_k \in M_k$, such that for all $k \ge 1$, either

(a) $x_k \in B_{g_k}(p_k, 200)$ and $|\mathrm{Rm}_k|(x_k) \ge k$, or

(b) $x_k \notin B_{g_k}(p_k, 200)$ and $\frac{|\mathrm{Rm}_k|}{f_k}(x_k) \ge k$.

Notice that by (2-5), we have $f_k(x) \ge 1000$ whenever $x \notin B(p_k, 200)$; hence $|\mathrm{Rm}_k|(x_k) \to \infty$.

The following standard point picking technique is due to Perelman [2002].

Claim 1. There exists $A_k \to \infty$ and $y_k \in B_{g_k}(x_k, 1)$, such that

$$|\mathrm{Rm}_k|(x) \le 2Q_k \quad \text{for all } x \in B_{g_k}(y_k, A_k Q_k^{-1/2}) \subset B_{g_k}(x_k, 2), \tag{4-1}$$

where $Q_k = |\mathrm{Rm}_k|(y_k) \ge |\mathrm{Rm}_k|(x_k)$.

Proof. Denote $Q_k^{(0)} = |\mathrm{Rm}_k|(x_k)$ and $y_k^{(0)} = x_k$, let $A_k = \frac{1}{100}(Q_k^{(0)})^{1/2} \to \infty$. We start from $y_k^{(0)}$. Suppose that $y_k^{(j)}$ is chosen and cannot be taken as y_k. Let $|\mathrm{Rm}_k|(y_k^{(j)}) = Q_k^{(j)}$. Then there exists $y_k^{(j+1)} \in B_{g_k}(y_k^{(j)}, A_k(Q_k^{(j)})^{-1/2})$, such that $Q_k^{(j+1)} = |\mathrm{Rm}_k|(y_k^{(j+1)}) \geq 2Q_k^{(j)}$. Hence we have

$$\begin{aligned}\mathrm{dist}_{g_k}(y_k^{(0)}, y_k^{(j+1)}) &\leq A_k(Q_k^{(0)})^{-1/2} + A_k(Q_k^{(1)})^{-1/2} + \cdots + A_k(Q_k^{(j)})^{-1/2} \\ &\leq A_k(Q_k^{(0)})^{-1/2}\left(1 + \frac{1}{\sqrt{2}} + \cdots + \left(\frac{1}{\sqrt{2}}\right)^j + \cdots\right) \\ &\leq \frac{1}{100} \times 4,\end{aligned}$$

and it follows that $y_k^{(j)} \in B_{g_k}(x_k, 1)$ for all $j \geq 0$. This procedure must terminate in finite steps since the manifold M_k is smooth; then the last element chosen by this procedure can be taken as y_k. □

Since for any $k \geq 1$ there can be only two cases (a) or (b), then either for infinitely many k, (a) holds, or, for infinitely many k, (b) holds. By passing to a subsequence, we need only to deal with the following two cases.

Case I. $x_k \in B_{g_k}(p_k, 200)$ and $|\mathrm{Rm}_k|(x_k) \geq k$, for all $k \geq 1$.

Case II. $x_k \notin B_{g_k}(p_k, 200)$ and $\frac{|\mathrm{Rm}_k|}{f_k}(x_k) \geq k$, for all $k \geq 1$.

We first consider Case I. We use Claim 1 to find $y_k \in B_{g_k}(x_k, 1)$, $Q_k = |\mathrm{Rm}_k|(y_k) \geq |\mathrm{Rm}_k|(x_k) \to \infty$, and $A_k \to \infty$ such that (4-1) holds. By (2-5) we have

$$R_k + |\nabla f_k|^2 = f_k \leq 10^5$$

on $B_{g_k}(y_k, A_k Q_k^{-1/2}) \subset B_{g_k}(p_k, 202)$. We scale g_k with the factor Q_k and use the notations with overlines to denote the scaled geometric quantities, that is, $\bar{g}_k = Q_k g_k$, $\overline{\mathrm{Rm}}_k = \mathrm{Rm}(\bar{g}_k)$, etc. Then we have that

$$\overline{\mathrm{Ric}}_k + \overline{\nabla}^2 f_k = \frac{Q_k^{-1}}{2}\bar{g}_k, \tag{4-2}$$

and that

$$|\overline{\mathrm{Rm}}_k| \leq 2, \tag{4-3}$$

$$\bar{R}_k + |\overline{\nabla} f_k|^2 \leq \frac{10^5}{Q_k} := r_k^2 \to 0, \tag{4-4}$$

on $B_{\bar{g}_k}(y_k, A_k)$, and by Proposition 3.1 and Proposition 3.2 that

$$|\overline{\nabla}\,\overline{\mathrm{Rm}}_k| \leq C_1, \tag{4-5}$$

$$|\overline{\mathrm{Ric}}_k| \leq C_2 r_k, \tag{4-6}$$

on $B_{\bar{g}_k}(y_k, A_k - 2)$, where C_1 is a constant depending only on the dimension $n = 4$, and C_2 is a constant depending only on the dimension $n = 4$ and $\kappa > 0$. We

can apply (4-3), (4-5), and the strong κ-noncollapsing assumption to extract from $\{(B_{\bar{g}_k}(y_k, A_k - 2), \bar{g}_k, y_k)\}_{k=1}^{\infty}$ a subsequence that converges in the pointed $C^{2,\alpha}$ Cheeger–Gromov sense to a complete nonflat Riemannian manifold $(M_\infty, g_\infty, y_\infty)$ with $|\mathrm{Rm}_\infty|(x_\infty) = 1$. By (4-6), (M_∞, g_∞) must be Ricci-flat and therefore has Euclidean volume growth, since it is also strongly κ-noncollapsed. By Corollary 8.86 of [Cheeger and Naber 2015], (M_∞, g_∞) is asymptotically locally Euclidean (ALE). By the definition of ALE, we have that outside a compact set M_∞ is diffeomorphic to a finite quotient of $\mathbb{R}^4 \setminus B(O, 1)$, it follows that there exists an open set $U_\infty \subset M_\infty$ containing the point y_∞, such that $\bar{U}_\infty$ is compact and that M_∞ is diffeomorphic to U_∞. By the definition of the pointed Cheeger–Gromov convergence, U_∞ can be embedded in infinitely many elements of the sequence $\{(M_k^4, f_k, g_k, p_k)\}_{k=1}^{\infty}$, and the images of the embeddings are bounded open sets. Furthermore, every one in the sequence of shrinkers satisfies (b) in Definition 1.1; it follows that U_∞ can also be embedded in a closed 4-manifold with vanishing second homology group, which is a contradiction against Theorem 6.1.

Case II is almost the same as Case I. By the same point picking and scaling method we also get (4-2), (4-3), (4-5), and (4-6). The only place where special care should be taken is (4-4). Notice that by (2-5), we have that $f_k(x) \geq 1000$ whenever $\mathrm{dist}_{g_k}(x, p_k) \geq 198$. Moreover, since $|\nabla\sqrt{f_k}| \leq \frac{1}{2}$, we have

$$\sqrt{f_k(x)} \leq \sqrt{f_k(x_k)} + 1 \leq \sqrt{\tfrac{10}{9} f_k(x_k)},$$

for all $x \in B_{g_k}(y_k, A_k Q_k^{-1/2}) \subset B_{g_k}(x_k, 2)$. It follows that

$$\bar{R}_k + |\overline{\nabla} f_k|^2 = \frac{f_k}{Q_k} \leq \frac{10}{9} \frac{f_k(x_k)}{|\mathrm{Rm}_k|(x_k)} := r_k^2 \to 0,$$

on $B_{\bar{g}_k}(y_k, A_k)$. Therefore (4-4) also holds in Case II and we obtain the same contradiction as in Case I. □

Theorem 4.2. *There exists $C < \infty$ depending only on κ, such that the following holds. Let $(M^4, g, f, p) \in \mathcal{N}^4(\kappa)$ be normalized as in* (1-2)*. Then it holds that*

$$|\mathrm{Rm}|(x) \leq C \quad \text{for all } x \in M.$$

Proof. We argue by contradiction. Suppose the statement is not true; then there exist a sequence of counterexamples $\{(M_k^4, f_k, g_k, p_k)\}_{k=1}^{\infty} \subset \mathcal{N}^4(\kappa)$ normalized as in (1-2), such that $\sup_{x \in M_k} |\mathrm{Rm}_k(x)| \to \infty$. By shifting the base points p_k and replacing f_k by $f_k - f_k(p_k)$, we may assume that for each k

$$Q_k := |\mathrm{Rm}_k(p_k)| \geq \tfrac{1}{2} \sup_{x \in M_k} |\mathrm{Rm}_k(x)| \to \infty.$$

Now we scale the sequence $\{(M_k^4, f_k, g_k, p_k)\}_{k=1}^{\infty}$ by the factors Q_k and use the notations with overlines to denote the scaled quantities as before, that is, $\bar{g}_k = Q_k g_k$,

$\overline{\mathrm{Rm}}_k = \mathrm{Rm}(\bar{g}_k)$, etc. Then we have

$$\overline{\mathrm{Ric}}_k + \overline{\nabla}^2 f_k = 0, \qquad |\overline{\nabla} f_k|^2 + \bar{R}_k = \frac{1}{Q_k} := r_k^2 \to 0,$$
$$|\overline{\mathrm{Rm}}_k|(x) \le 2 \quad \text{for all } x \in M, \qquad |\overline{\mathrm{Rm}}_k|(p_k) = 1.$$

Recall that all $(M_k, \bar{g}_k)$ are κ-noncollapsed on call scales with respect to a uniform $\kappa > 0$. It then follows from Propositions 3.1 and 3.2 that there exists $C < \infty$ and $C_l < \infty$ for each $l \in \mathbb{Z}_+$, where C, κ and C_l's depend only on the dimension, such that

$$|\overline{\nabla}^l \overline{\mathrm{Rm}}_k| \le C_l, \quad |\overline{\mathrm{Ric}}_k| \le C r_k \to 0.$$

Hence, by the noncollapsing condition, we can extract from $\{(M_k, \bar{g}_k, p_k)\}_{k=1}^{\infty}$ a subsequence that converges in the smooth pointed Cheeger–Gromov sense to a smooth manifold $(M_\infty, g_\infty, p_\infty)$. By the choice of p_k's, we have that $|\mathrm{Rm}_\infty|(p_\infty) = 1 > 0$, hence g_∞ is nonflat. Since $|\overline{\mathrm{Ric}}_k|$ converges to 0 uniformly, we have that g_∞ is Ricci flat. Finally, since (M_∞, g_∞) is also strongly κ-noncollapsed on all scales, it has Euclidean volume growth, and hence must be ALE by Corollary 8.86 in [Cheeger and Naber 2015]. The rest of the proof now follows similarly from Theorem 4.1. □

5. Proof of the main theorems

Proof of Theorem 1.4. By Theorem 4.1, Proposition 3.1, and (2-5), we obtain locally uniform bounds for the curvatures, the derivatives of the curvatures, and the potential functions for any sequence in the space $\mathcal{M}^4(\kappa)$. Applying the standard regularity theorem to the elliptic equation $\Delta f = \frac{n}{2} - R$, we also obtain locally uniform bounds for the derivatives of the potential functions. Hence we can extract from any sequence contained in $\mathcal{M}^4(\kappa)$ a subsequence that converges in the smooth pointed Cheeger–Gromov sense to a shrinking gradient Ricci soliton, also normalized as in (1-1). It remains to show that the limit Ricci shrinker is in $\mathcal{M}^4(\kappa)$. Item (c) in Definition 1.1 is obvious, we proceed to show (a) and (b).

We let $\{(M_k, g_k, f_k, p_k)\}_{k=1}^{\infty} \subset \mathcal{M}^4(\kappa)$, all normalized as in (1-1), and we let $(M_\infty, g_\infty, f_\infty, p_\infty)$ be their limit Ricci shrinker in the smooth pointed Cheeger–Gromov sense, also normalized as in (1-1). By the definition of Cheeger–Gromov convergence, we have that every open bounded subset in (M_∞, g_∞) can be embedded in infinitely many (M_k, g_k)'s in the sequence, and the images of these embeddings are also bounded open sets, and therefore can be embedded in closed 4-manifolds with vanishing second homology group. To show that (M_∞, g_∞) is nonflat, we make the following observation.

Claim 2. $$\mathrm{Vol}_f(g_\infty) = \lim_{k\to\infty} \mathrm{Vol}_f(g_k),$$
where Vol_f is the f-volume defined by $\mathrm{Vol}_f(g) = \int_M e^{-f}\, dg$.

Proof. By the uniform rapid decay of e^{-f} (2-5) and the uniform volume growth bound (2-6) we have that for any $\eta > 0$, there exists $A_0 < \infty$ such that for all $A > A_0$ it holds that

$$\mathrm{Vol}_f(g_k) - \eta < \int_{B_{g_k}(p_k, A)} e^{-f_k}\, dg_k \le \mathrm{Vol}_f(g_k),$$

for every $k \ge 1$. The conclusion follows from first taking $k \to \infty$, and then $A \to \infty$, and finally $\eta \to 0$. □

By Proposition 2.5 and Claim 2 we have

$$\mathrm{Vol}_f(g_\infty) = \lim_{k\to\infty} \mathrm{Vol}_f(g_k) \le (4\pi)^{n/2}(1-\varepsilon),$$

where $\varepsilon > 0$ is given by Proposition 2.5. Hence $(M_\infty, g_\infty, f_\infty, p_\infty)$ is not flat, because the f-volume of the Gaussian shrinker is $(4\pi)^{n/2}$. This completes the proof of Theorem 1.4. □

Proof of Theorem 1.5. Combining Theorem 4.2, the fact that $|\nabla f| \le 1$ by (1-2), and Proposition 3.1, we have that any curvature derivative is uniformly bounded for all elements in $\mathcal{N}^4(\kappa)$. Furthermore, since

$$f(p) = 0, \quad |\nabla f| \le 1, \quad |\nabla^2 f| = |\mathrm{Ric}| \le C(\kappa),$$

we have a uniform growth estimate for $|f|$, and we can derive uniform higher derivative estimates for f by using the elliptic equation $\Delta f = -R$. Taking into account the noncollapsing condition, we immediately obtain the precompactness. By the same argument as in the proof of Theorem 1.4, we have that every possible limit of a convergent sequence in $\mathcal{N}^4(\kappa)$ must satisfy (b) and (c) in Definition 1.2. Such a limit can be either nonflat, hence lies in $\mathcal{N}^4(\kappa)$, or, flat, hence must be the Euclidean space because of its maximum volume growth by (c). This completes the proof of the theorem. □

Proof of Corollary 1.6. To prove (a) we argue by contradiction. Suppose there exists $\{(M_k, g_k, f_k, p_k)\}_{k=1}^\infty \subset \mathcal{M}^4(\kappa)$ such that $R_k(p_k) \to 0$. By Theorem 1.4 we can extract a subsequence that converges to a shrinking gradient Ricci soliton $(M_\infty, g_\infty, f_\infty, p_\infty)$ with $R_\infty(p_\infty) = 0$, which by Chen [2009] is flat; this is a contradiction.

To prove (b), we recall that by the proof of Chow, Lu, and Yang [Chow et al. 2011], we only need a uniform upper bound for f and a uniform lower bound for R on a sufficiently large ball, say $B(p, 1000)$, where the former is given by (2-5) and the latter is proved in the same way as for (a).

To prove (c), we claim that there exist $c > 0$, depending only on κ, such that $\mathrm{Vol}_f(g) > c$ for all $(M, g, f, p) \in \mathcal{M}^4(\kappa)$. Suppose this is not true. As in the proof of (a), we can find a sequence of counterexamples converging to a Ricci shrinker $(M_\infty, g_\infty, f_\infty, p_\infty)$ with $\mathrm{Vol}_f(g_\infty) = 0$, which is a contradiction. Hence we have

$\mathrm{Vol}_f(g) \in [c, (4\pi)^{n/2}(1-\varepsilon)]$ for all $(M, g, f, p) \in \mathcal{M}^4(\kappa)$. The conclusion follows from [Wylie 2008] and [Chow and Lu 2016]. □

6. Excluding instantons by a topological condition

In this section we provide an alternative proof for Corollary 5.8 of [Anderson 2010]. This proof is based in essence altogether on the personal notes of Richard Bamler, to whom we are indebted for graciously providing them. However, any mistakes in transcription is solely due to the author. Forasmuch as Anderson's result is of fundamental importance to our main theorem, we include this section for the sake of completeness to help the readers to follow some details.

Theorem 6.1. *Let N be a smooth closed 4-dimensional manifold such that*

$$H_2(N) = 0, \tag{6-1}$$

where H_2 is the second homology group with coefficients in $\mathbb{Z}$. Then there is no open subset $U \subset N$ with the property that U admits an Einstein ALE metric.

We split the proof into the following lemmas.

Lemma 6.2. *Let N be the closed manifold in the statement of Theorem 6.1. Let $U \subset N$ be an connected open subset such that $\partial U \cong \mathbb{S}^3/\Gamma$ and $H_1(U, \partial U) = 0$, where Γ is a finite group. Then the following hold.*

$$H_1(\partial U) = H_1(U) \oplus H_1(U), \tag{6-2}$$

$$H_2(U) = 0. \tag{6-3}$$

Proof. By Poincaré duality, we have

$$H^2(N; \mathbb{Z}) \cong H_2(N) = 0.$$

By the universal coefficient theorem, we have

$$0 = H^2(N; \mathbb{Z}) \cong \mathrm{Hom}(H_2(N), \mathbb{Z}) \oplus \mathrm{Ext}(H_1(N), \mathbb{Z}),$$

which implies that $H_1(N)$ is torsion free, so henceforth we may assume

$$H_1(N) \cong \mathbb{Z}^d, \tag{6-4}$$

where $d \geq 0$. By Poincaré duality and by the universal coefficient theorem, we have

$$H_2(\partial U) \cong H^1(\partial U) \cong \mathrm{Hom}(H_1(\partial U), \mathbb{Z}) = 0,$$

where the last equality is because $H_1(\partial U)$ is a finite abelian group and hence purely torsion. Then we have the Mayer–Vietoris sequence

$$0 = H_2(\partial U) \to H_2(U) \oplus H_2(N \backslash U) \to H_2(N) = 0,$$

from whence we obtain

$$H_2(U) = H_2(N \backslash U) = 0. \tag{6-5}$$

On the other hand, we consider the long exact sequence

$$0 = H_2(U) \to H_2(U, \partial U) \to H_1(\partial U) \to H_1(U) \to H_1(U, \partial U) = 0. \tag{6-6}$$

It follows that $H_1(U)$ is purely torsion since the third homomorphism above is surjective and $H_1(\partial U)$ is finite. Hence by the universal coefficient theorem and by Poincaré–Lefschetz duality we have

$$H_2(U, \partial U) \cong H^2(U) \cong \mathrm{Hom}(H_2(U), \mathbb{Z}) \oplus \mathrm{Ext}(H_1(U), \mathbb{Z}) \cong H_1(U),$$

where the last isomorphism follows from the fact that $H_1(U)$ is purely torsion and (6-5). Hence (6-6) is simplified as

$$0 \to H_1(U) \to H_1(\partial U) \to H_1(U) \to 0. \tag{6-7}$$

To see (6-7) splits, we consider the following Mayer–Vietoris sequence

$$0 = H_2(N) \to H_1(\partial U) \to H_1(U) \oplus H_1(N \backslash U) \to H_1(N) \cong \mathbb{Z}^d,$$

where we have used (6-4). If we write $H_1(U) \oplus H_1(N\backslash U) \cong H_1(U) \oplus T \oplus \mathbb{Z}^e$, where T is the torsion part of $H_1(N\backslash U)$, since $H_1(\partial U)$ is purely torsion, we have that the image of the second homomorphism in the above sequence is in $H_1(U) \oplus T \oplus \{0\}$, whose image under the third homomorphism is 0. Hence we can simplify the above sequence as

$$0 \to H_1(\partial U) \to H_1(U) \oplus T \to 0.$$

The inclusion $H_1(U) \hookrightarrow H_1(U) \oplus T \cong H_1(\partial U)$ gives a homomorphism $H_1(U) \to H_1(\partial U)$, whose composition with the third homomorphism in (6-7) is the identity on $H_1(U)$. It follows that (6-7) splits and we have completed the proof. □

Lemma 6.3. *Let $\mathbb{S}^3/\Gamma$ be a round space form, where Γ is a finite group. If $H_1(\mathbb{S}^3/\Gamma) \cong G \oplus G$, for some group G, then either Γ is the binary dihedral group D_n^* with n being even, or Γ is the binary icosahedral group with order 120.*

Proof. We shall check every possible group Γ.

(a) **Lens space:** In this case $H_1(\mathbb{S}^3/\Gamma) \cong \Gamma = \mathbb{Z}_m$ with $m \geq 2$, which is not possible.

(b) **Prism manifold:** In this case the fundamental group has the presentation

$$\langle x, y \mid xyx^{-1} = y^{-1},\ x^{2^k} = y^n \rangle \times \mathbb{Z}_m,$$

where $k, m \geq 1$, $n \geq 2$, and m is coprime to $2n$. Its abelianization is

$$H_1(\mathbb{S}^3/\Gamma) \cong \langle x, y \mid y = y^{-1},\ x^{2^k} = y^n \rangle \times \mathbb{Z}_m,$$

where $y^2 = 1$. We have that $H_1(\mathbb{S}^3/\Gamma) \cong \mathbb{Z}_2 \times \mathbb{Z}_{2^k} \times \mathbb{Z}_m$ in the case n is even, and that $H_1(\mathbb{S}^3/\Gamma) \cong \mathbb{Z}_{2^{k+1}} \times \mathbb{Z}_m$ in the case n is odd. Since m is coprime to $2n$, we have that the only possible case is when $m = 1$, n is even, and $k = 1$.

(c) **Tetrahedral manifold:** In this case we have

$$\Gamma = \langle x, y, z \mid (xy)^2 = x^2 = y^2,\ zxz^{-1} = y,\ zyz^{-1} = xy,\ z^{3^k} = 1\rangle \times \mathbb{Z}_m,$$

where $k, m \geq 1$ and m is coprime to 6. Then we have

$$\begin{aligned} H_1(\mathbb{S}^3/\Gamma) &\cong \langle x, y, z \mid x^2 = y^2 = 1,\ x = y,\ y = xy,\ z^{3^k} = 1\rangle \times \mathbb{Z}_m \\ &= \langle x, z \mid x^2 = 1,\ x = x^2,\ z^{3^k} = 1\rangle \times \mathbb{Z}_m = \mathbb{Z}_{3^k} \times \mathbb{Z}_m. \end{aligned}$$

Since m is coprime to 6, this case is not possible.

(d) **Octahedral manifold:** In this case we have

$$\Gamma = \langle x, y \mid (xy)^2 = x^3 = y^4\rangle \times \mathbb{Z}_m,$$

where m is coprime to 6. Then we have

$$H_1(\mathbb{S}^3/\Gamma) \cong \langle x, y \mid x = y^2 = x^2\rangle \times \mathbb{Z}_m = \mathbb{Z}_2 \times \mathbb{Z}_m.$$

Since m is coprime to 6, this case is not possible.

(e) **Icosahedral manifold:** In this case we have

$$\Gamma = \langle x, y \mid (xy)^2 = x^3 = x^3y^5\rangle \times \mathbb{Z}_m,$$

where m is coprime to 30. Then we have

$$\begin{aligned} H_1(\mathbb{S}^3/\Gamma) &\cong \langle x, y \mid x = y^2,\ x^2 = y^3\rangle \times \mathbb{Z}_m \\ &= \langle x, y \mid x = y^2,\ x^2 = y^3,\ y = 1\rangle \times \mathbb{Z}_m = \mathbb{Z}_m. \end{aligned}$$

The only possibility is $m = 1$. □

We still need to consider the two cases when Γ is the binary dihedral group D^*_{2n} or the binary icosahedral group. In both cases Γ can be embedded in $SU(2)$. Indeed, it is well known that the binary dihedral, tetrahedral, octahedral, and icosahedral groups are all finite subgroups of $SU(2)$; see [Kronheimer 1989].

Lemma 6.4. *Let $\mathbb{S}^3/\Gamma$ be the spherical space form with $\Gamma < O(4)$ being either the binary dihedral group D^*_{2n} or the binary icosahedral group. Then there exists a complex structure on $\mathbb{R}^4$ such that $\Gamma < SU(2)$*

Lemma 6.5. *Let (U, g) be an Einstein ALE space which is asymptotic to $\mathbb{S}^3/\Gamma$, where $\Gamma < SU(2)$ is isomorphic to the binary dihedral group D^*_{2n} or to the binary icosahedral group. Then $b_2(U) \geq 1$.*

Proof. Assume that $b_2(U) = 0$. Then we have $\chi(U) = 1 - b_1(U) - b_3(U) \leq 1$ and $\tau(U) = 0$. Using the Chern–Gauss–Bonnet theorem and the Atiyah–Patodi–Singer

index theorem, we have (see (4.4) and (4.5) in [Nakajima 1990])

$$1 \geq \chi(U) = \frac{1}{8\pi^2}\int_U |W|^2\, dg + \frac{1}{|\Gamma|},$$
$$0 = \tau(U) = \frac{1}{12\pi^2}\int_U (|W^+|^2 - |W^-|^2)\, dg - \eta_S(\mathbb{S}^3/\Gamma),$$

where η_S stands for the *eta invariant*. Hence we have

$$\frac{2}{3} \geq \frac{2}{12\pi^2}\int_U |W^-|^2\, dg + \frac{2}{3|\Gamma|} + \eta_S(\mathbb{S}^3/\Gamma),$$

which implies

$$\eta_S(\mathbb{S}^3/\Gamma) \leq \frac{2}{3}\Big(1 - \frac{1}{|\Gamma|}\Big) < \frac{2}{3}.$$

On the other hand, by [Nakajima 1990], if Γ is the binary dihedral group D^*_{2n}, we have

$$\eta_S(\mathbb{S}^3/\Gamma) = \frac{2(2n+2)^2 - 8(2n+2) + 9}{6\cdot 2n} = \frac{8n^2+1}{12n} > \frac{2}{3}.$$

Similarly, if Γ is the binary icosahedral group, then we have

$$\eta_S(\mathbb{S}^3/\Gamma) = \frac{361}{180} > \frac{2}{3}.$$

In either case we yield a contradiction. □

Proof of Theorem 6.1. Let N be the manifold in Theorem 6.1 and $U \subset N$ be a connected open subset that admits an Einstein ALE metric.

Claim 3. $$H_1(U, \partial U) = 0.$$

Proof. Suppose the claim does not hold. We first show that the boundary $\partial\widetilde{U}$ of the universal cover $\widetilde{U}$ has more than one component. Since ∂U is a deformation retraction of its collar neighbourhood, by excision we have

$$H_1(U/\partial U, \partial U/\partial U) = H_1(U, \partial U) \neq 0.$$

Hence we have $\pi_1(U/\partial U) \neq 0$. Let γ_0 be a loop in $U/\partial U$ based at $\partial U/\partial U$ that is not null-homotopic. Lifting this loop to U by the quotient map $q : U \to U/\partial U$, we obtain a curve γ in U, whose ends lie in ∂U. By using the universal covering map $p : \widetilde{U} \to U$, we can lift γ to $\widetilde{\gamma}$, a curve in $\widetilde{U}$ whose ends lie in $\partial\widetilde{U}$. If $\partial\widetilde{U}$ is connected, since $\widetilde{U}$ is simply connected, we have that $\widetilde{\gamma}$ is homotopic to a curve that lies in $\partial\widetilde{U}$. Composing this homotopy with $q \circ p$ we obtain a homotopy between γ_0 with a point; this is a contradiction. Hence $\partial\widetilde{U}$ has more than one component.

Next we observe that if U admits an Einstein ALE metric, then we can lift this metric to $\widetilde{U}$, which has more than one end. By Cheeger–Gromoll's splitting theorem, $\widetilde{U}$ splits as the product of a line and a Ricci flat 3-manifold; hence this metric is flat, which is a contradiction. □

We continue the proof of Theorem 6.1. By (6-3) we have that $b_2(U) = 0$. On the other hand, combining (6-2) and Lemma 6.3 we have that $\partial U \cong \mathbb{S}^3/\Gamma$, where Γ is either the binary dihedral group D^*_{2n} or the binary icosahedral group. It follows from Lemma 6.5 that $b_2(U) \geq 1$, and we obtain a contradiction. □

Acknowledgements

The author would like to thank his doctoral advisors, Professor Bennett Chow and Professor Lei Ni, for their constant help. He also owes many thanks to Professor Richard Bamler, who provided him with the personal notes appearing in Section 6. The idea of this paper was inspired by a discussion between Professor Bennett Chow and the author.

References

[Anderson 2010] M. T. Anderson, "A survey of Einstein metrics on 4-manifolds", pp. 1–39 in *Handbook of geometric analysis*, vol. 3, edited by L. Ji et al., Adv. Lect. Math. (ALM) **14**, International Press, Somerville, MA, 2010. MR Zbl

[Appleton 2017] A. Appleton, "A family of non-collapsed steady Ricci solitons in even dimensions greater or equal to four", preprint, 2017. arXiv

[Bamler and Zhang 2017] R. H. Bamler and Q. S. Zhang, "Heat kernel and curvature bounds in Ricci flows with bounded scalar curvature", *Adv. Math.* **319** (2017), 396–450. MR Zbl

[Brendle 2013] S. Brendle, "Rotational symmetry of self-similar solutions to the Ricci flow", *Invent. Math.* **194**:3 (2013), 731–764. MR Zbl

[Bryant 2005] R. L. Bryant, "Ricci flow solitons in dimension three with SO(3)-symmetries", preprint, Duke University, 2005, https://services.math.duke.edu/∼bryant/3DRotSymRicciSolitons.pdf.

[Cao and Zhou 2010] H.-D. Cao and D. Zhou, "On complete gradient shrinking Ricci solitons", *J. Differential Geom.* **85**:2 (2010), 175–185. MR Zbl

[Carrillo and Ni 2009] J. A. Carrillo and L. Ni, "Sharp logarithmic Sobolev inequalities on gradient solitons and applications", *Comm. Anal. Geom.* **17**:4 (2009), 721–753. MR Zbl

[Cheeger and Naber 2015] J. Cheeger and A. Naber, "Regularity of Einstein manifolds and the codimension 4 conjecture", *Ann. of Math.* (2) **182**:3 (2015), 1093–1165. MR Zbl

[Chen 2009] B.-L. Chen, "Strong uniqueness of the Ricci flow", *J. Differential Geom.* **82**:2 (2009), 363–382. MR Zbl

[Chow and Lu 2016] B. Chow and P. Lu, "A bound for the order of the fundamental group of a complete noncompact Ricci shrinker", *Proc. Amer. Math. Soc.* **144**:6 (2016), 2623–2625. MR Zbl

[Chow et al. 2006] B. Chow, P. Lu, and L. Ni, *Hamilton's Ricci flow*, Graduate Studies in Mathematics **77**, American Mathematical Society, Providence, RI, 2006. MR Zbl

[Chow et al. 2007] B. Chow, S.-C. Chu, D. Glickenstein, C. Guenther, J. Isenberg, T. Ivey, D. Knopf, P. Lu, F. Luo, and L. Ni, *The Ricci flow: techniques and applications, Part I: Geometric aspects*, Mathematical Surveys and Monographs **135**, American Mathematical Society, Providence, RI, 2007. MR Zbl

[Chow et al. 2011] B. Chow, P. Lu, and B. Yang, "Lower bounds for the scalar curvatures of noncompact gradient Ricci solitons", *C. R. Math. Acad. Sci. Paris* **349**:23-24 (2011), 1265–1267. MR Zbl

[Croke 1980] C. B. Croke, "Some isoperimetric inequalities and eigenvalue estimates", *Ann. Sci. École Norm. Sup.* (4) **13**:4 (1980), 419–435. MR Zbl

[Enders et al. 2011] J. Enders, R. Müller, and P. M. Topping, "On type-I singularities in Ricci flow", *Comm. Anal. Geom.* **19**:5 (2011), 905–922. MR Zbl

[Feldman et al. 2003] M. Feldman, T. Ilmanen, and D. Knopf, "Rotationally symmetric shrinking and expanding gradient Kähler–Ricci solitons", *J. Differential Geom.* **65**:2 (2003), 169–209. MR Zbl

[Gu and Zhu 2008] H.-L. Gu and X.-P. Zhu, "The existence of type II singularities for the Ricci flow on S^{n+1}", *Comm. Anal. Geom.* **16**:3 (2008), 467–494. MR Zbl

[Hamilton 1982] R. S. Hamilton, "Three-manifolds with positive Ricci curvature", *J. Differential Geom.* **17**:2 (1982), 255–306. MR Zbl

[Hamilton 1993] R. S. Hamilton, "Eternal solutions to the Ricci flow", *J. Differential Geom.* **38**:1 (1993), 1–11. MR Zbl

[Hamilton 1995a] R. S. Hamilton, "A compactness property for solutions of the Ricci flow", *Amer. J. Math.* **117**:3 (1995), 545–572. MR Zbl

[Hamilton 1995b] R. S. Hamilton, "The formation of singularities in the Ricci flow", pp. 7–136 in *Surveys in differential geometry* (Cambridge, MA, 1993), vol. II, edited by S.-T. Yau, Int. Press, Cambridge, MA, 1995. MR Zbl

[Haslhofer and Müller 2011] R. Haslhofer and R. Müller, "A compactness theorem for complete Ricci shrinkers", *Geom. Funct. Anal.* **21**:5 (2011), 1091–1116. MR Zbl

[Haslhofer and Müller 2015] R. Haslhofer and R. Müller, "A note on the compactness theorem for 4d Ricci shrinkers", *Proc. Amer. Math. Soc.* **143**:10 (2015), 4433–4437. MR Zbl

[Kronheimer 1989] P. B. Kronheimer, "The construction of ALE spaces as hyper-Kähler quotients", *J. Differential Geom.* **29**:3 (1989), 665–683. MR Zbl

[Li et al. 2018] X. Li, L. Ni, and K. Wang, "Four-dimensional gradient shrinking solitons with positive isotropic curvature", *Int. Math. Res. Not.* **2018**:3 (2018), 949–959. MR Zbl

[Munteanu 2009] O. Munteanu, "The volume growth of complete gradient shrinking Ricci solitons", preprint, 2009. arXiv

[Munteanu and Wang 2011] O. Munteanu and J. Wang, "Smooth metric measure spaces with non-negative curvature", *Comm. Anal. Geom.* **19**:3 (2011), 451–486. MR Zbl

[Munteanu and Wang 2012] O. Munteanu and J. Wang, "Analysis of weighted Laplacian and applications to Ricci solitons", *Comm. Anal. Geom.* **20**:1 (2012), 55–94. MR Zbl

[Munteanu and Wang 2014] O. Munteanu and J. Wang, "Geometry of manifolds with densities", *Adv. Math.* **259** (2014), 269–305. MR Zbl

[Munteanu and Wang 2015] O. Munteanu and J. Wang, "Geometry of shrinking Ricci solitons", *Compos. Math.* **151**:12 (2015), 2273–2300. MR Zbl

[Munteanu and Wang 2016] O. Munteanu and J. Wang, "Structure at infinity for shrinking Ricci solitons", preprint, 2016. arXiv

[Munteanu and Wang 2017a] O. Munteanu and J. Wang, "Conical structure for shrinking Ricci solitons", *J. Eur. Math. Soc.* (*JEMS*) **19**:11 (2017), 3377–3390. MR Zbl

[Munteanu and Wang 2017b] O. Munteanu and J. Wang, "Positively curved shrinking Ricci solitons are compact", *J. Differential Geom.* **106**:3 (2017), 499–505. MR Zbl

[Naber 2010] A. Naber, "Noncompact shrinking four solitons with nonnegative curvature", *J. Reine Angew. Math.* **645** (2010), 125–153. MR Zbl

[Nakajima 1990] H. Nakajima, "Self-duality of ALE Ricci-flat 4-manifolds and positive mass theorem", pp. 385–396 in *Recent topics in differential and analytic geometry*, edited by T. Ochiai, Adv. Stud. Pure Math. **18**, Academic Press, Boston, 1990. MR Zbl

[Perelman 2002] G. Perelman, "The entropy formula for the Ricci flow and its geometric applications", preprint, 2002. Zbl arXiv

[Shi 1989] W.-X. Shi, "Deforming the metric on complete Riemannian manifolds", *J. Differential Geom.* **30**:1 (1989), 223–301. MR Zbl

[Wang 2012] B. Wang, "On the conditions to extend Ricci flow (II)", *Int. Math. Res. Not.* **2012**:14 (2012), 3192–3223. MR Zbl

[Wylie 2008] W. Wylie, "Complete shrinking Ricci solitons have finite fundamental group", *Proc. Amer. Math. Soc.* **136**:5 (2008), 1803–1806. MR Zbl

[Yokota 2009] T. Yokota, "Perelman's reduced volume and a gap theorem for the Ricci flow", *Comm. Anal. Geom.* **17**:2 (2009), 227–263. MR Zbl

[Yokota 2012] T. Yokota, "Addendum to 'Perelman's reduced volume and a gap theorem for the Ricci flow"', *Comm. Anal. Geom.* **20**:5 (2012), 949–955. MR Zbl

Received August 16, 2018. Revised January 15, 2019.

YONGJIA ZHANG
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MINNESOTA
TWIN CITIES, MN 55414
UNITED STATES
zhan7298@umn.edu

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 303 No. 1 November 2019

0030-8730(201911)303:1;1-J