Pacific Journal of Mathematics

Volume 266 No. 1 November 2013

PACIFIC JOURNAL OF MATHEMATICS

msp.org/pjm

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

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

The subscription price for 2013 is US \$400/year for the electronic version, and \$485/year for print and electronic.
Subscriptions, requests for back issues and changes of subscribers 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 the Science Citation Index.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.1

MULTIPLICITY OF SOLUTIONS TO THE YAMABE PROBLEM ON COLLAPSING RIEMANNIAN SUBMERSIONS

RENATO G. BETTIOL AND PAOLO PICCIONE

Let g_t be a family of constant scalar curvature metrics on the total space of a Riemannian submersion obtained by shrinking the fibers of an original metric g, so that the submersion *collapses* as $t \to 0$ (that is, the total space converges to the base in the Gromov–Hausdorff sense). We prove that, under certain conditions, there are at least 3 unit volume constant scalar curvature metrics in the conformal class $[g_t]$ for infinitely many t accumulating at 0. This holds, for instance, for homogeneous metrics g_t obtained via Cheeger deformation of homogeneous fibrations with fibers of positive scalar curvature.

1. Introduction

A classic problem in Riemannian geometry is to find possible *canonical* metrics on a given smooth manifold M. Along this quest, an important achievement was the complete solution of the celebrated Yamabe problem, which states that given a closed Riemannian manifold (M, g_0) with $\dim M \geq 3$, there exists a constant scalar curvature metric g conformal to g_0. Such a metric g can be characterized variationally as a critical point of the Hilbert–Einstein functional

$$\mathscr{A}(g) = \frac{1}{\operatorname{Vol}(g)} \int_M \operatorname{scal}(g) \operatorname{vol}_g, \tag{1-1}$$

restricted to the set $[g_0]_1$ of unit volume metrics in the conformal class of g_0. Existence of a metric that *minimizes* this constrained functional, called the *Yamabe metric*, is a consequence of the works of Yamabe [1960], Trudinger [1968], Aubin [1976] and Schoen [1984]. In addition to the minimizer, there may also be other critical points; thus the solution *may be not unique*. However, Anderson [2005]

The second named author is supported by FAPESP, São Paulo, Brazil, and by CNPq, Brazil. This paper was concluded during Bettiol's visit to the University of São Paulo in May 2012. We thankfully acknowledge the financial support for this visit from FAPESP, processo 2011/21362-2, "Group actions, submanifold theory and global analysis in Riemannian and pseudo-Riemannian geometry".
MSC2010: primary 53C30, 58J55; secondary 53A30, 53C20, 53C21, 58E50, 58J50.
Keywords: Yamabe problem, constant scalar curvature metrics, equivariant bifurcation, Riemannian submersions, Cheeger deformation, homogeneous fibration.

recently proved that, on generic conformal classes, the Yamabe metric is the unique solution. In the present paper, we are interested in the complementary situation, that is, finding conformal classes where the Yamabe problem has *multiple* solutions. Our main results provide a large class of manifolds whose conformal class contains at least 3 distinct solutions (see Theorems A and B).

A classic method to obtain new solutions of a PDE from a path of known solutions is to use *bifurcation theory*. The basic setup for our framework consists of a given one-parameter family g_t of known solutions to the Yamabe problem,

$$\mathrm{d}\big(\mathscr{A}|_{[g_t]_1}\big)(\hat{g}_t) = 0 \quad \text{for } t \in [a, b], \tag{1-2}$$

where $\hat{g}_t$ is the unit volume metric homothetic to g_t. We study the case where g_t is obtained by shrinking the fibers of a Riemannian submersion with totally geodesic fibers. By proving that certain topological invariants (for example, the Morse index) of g_t change as t crosses a value t_*, one obtains existence of *new* solutions accumulating at g_{t_*}. Then, a simple trick (Proposition 7.1) implies that for t's close to t_* there are at least 3 solutions to the Yamabe problem on $[g_t]_1$.

In our main result, g_t are homogeneous metrics, hence trivially solutions to the Yamabe problem and good (that is, nongeneric) candidates for admitting other solutions in their conformal class. Let $H \subsetneq K \subsetneq G$ be compact connected Lie groups with $\dim K/H \geq 2$, and assume that either H is normal in K or K is normal in G. Consider the *homogeneous fibration*

$$K/H \to G/H \xrightarrow{\pi} G/K, \quad \pi(gH) = gK. \tag{1-3}$$

More precisely, π is the associated bundle with fiber K/H to the K-principal bundle $G \to G/H$. Endow the above spaces with *compatible* homogeneous metrics (see Section 4). Shrinking the fibers of (1-3) by a factor t^2, we get a family g_t of homogeneous metrics on G/H, sometimes called *canonical variation* of (1-3). This is, up to reparametrization, the *Cheeger deformation* of G/H in the direction of the natural K-action. As t approaches 0, the manifolds $(G/H, g_t)$ converge to the base G/K in the Gromov–Hausdorff sense. We explore the existence of infinitely many bifurcations of (1-2) along this collapse of $(G/H, g_t)$ onto the base to obtain the following multiplicity result.

Theorem A. *Let $K/H \to G/H \to G/K$ be a homogeneous fibration endowed with compatible homogeneous metrics such that K/H has positive scalar curvature. Let g_t be the family of G-invariant metrics on G/H obtained as described above. Then, there exists a subset $\mathfrak{T} \subset\,]0, 1[$, accumulating at 0, such that for each $t \in \mathfrak{T}$ there are at least 2 solutions to the Yamabe problem in $[g_t]_1$, other than $\hat{g}_t$, and they are not G-invariant.*

Theorem A applies, for example, to short exact sequences $K \to G \to G/K$ of compact connected Lie groups and to *twisted product* fibrations

$$K/H \to ((K \times L)/\Gamma)/H \to G/K,$$

where K/H is a compact homogenous space of positive scalar curvature and $G = (K \times L)/\Gamma$ is a connected compact Lie group. More interestingly, the result also holds, for example, for any homogeneous fibrations (1-3) where H is normal in K and the quotient K/H is a nonabelian compact connected Lie group. In this case, G can have arbitrarily large rank and dimension, and the corresponding possible total spaces can be much more general than twisted products. For more details and examples, see Section 5. We also stress that the hypotheses made on the scalar curvature (and dimension) of the fibers K/H are necessary. A counterexample is given in the end of Section 4.

Various other nonuniqueness phenomena have been studied in the literature, but usually this can only be achieved for very specific examples; see [Ambrosetti and Malchiodi 1999; Berti and Malchiodi 2001; Bettiol and Piccione 2013; Fischer and Marsden 1975; Lima et al. 2012; Schoen 1991]. One exception is a remarkable result of Pollack [1993] that proved existence of arbitrarily C^0-small perturbations of any given metric, with an arbitrarily large number of solutions in its conformal class. Previously, Schoen [1991] had proven the existence of an increasing number of solutions, with larger energy and Morse index, in the conformal class of the product $S^1(r) \times S^{m-1}$ of round spheres, as r tends to infinity. Lima, Piccione and Zedda [Lima et al. 2012] generalized this result to families of product metrics on a product $M_1 \times M_2$ of compact Riemannian manifolds given by rescaling one of the factors, obtaining *bifurcation* of solutions. Inspired by these results, the authors recently obtained similar bifurcation results for families of homogeneous spheres in [Bettiol and Piccione 2013]. The core of Theorem A is a further generalization, establishing that such bifurcations indeed occur on several other families of compact homogeneous spaces.

The initial approach to detect bifurcation in a variational problem, such as (1-2), is to look for a change in the Morse index. This is done by identifying eigenvalues of the second variation (2-3) of $\mathcal{A}|_{[g_t]_1}$ that change sign for certain values of t. Nevertheless, a subtle *compensation* problem may occur when other eigenvalues with the same combined multiplicity cross zero in the opposite direction. On the one hand, it is not hard to detect passage through zero of certain explicitly computable parts of the spectrum. On the other hand, it is in general not possible to rule out that other eigenvalues also change sign at the same time, potentially producing this compensation that leaves the Morse index unchanged.

In the case of the Yamabe problem, the spectral analysis required consists of comparing eigenvalues of the Laplacian of Riemannian submersions with the scalar curvature. Since the fibers are assumed totally geodesic, eigenfunctions of the Laplacian of the base may be lifted to eigenfunctions of the total space that are constant along the fibers (see Section 3). This provides an easier subset of the spectrum to deal with, in the sense that a direct computation of the scalar curvature immediately gives infinitely many crossings through zero. Nevertheless, no information is available in general regarding possible compensation due to other crossings in the opposite direction.

The key to handle this situation is to use homogeneity, placing the problem in an equivariant context where a subtler bifurcation criterion due to Smoller and Wasserman [1990] applies. Linearizing the action, one gets a representation of the symmetry group in each eigenspace of the second variation. The direct sum of those representations that correspond to negative eigenvalues is called the *negative isotropic representation*; see page 8. The equivariant criterion asserts that bifurcation occurs at the degeneracy values where the negative isotropic representation changes, which is the case of all degeneracy values mentioned above that correspond to crossings of eigenvalues of the base. It is then a simple observation that such bifurcation yields the desired multiplicity result (see Proposition 7.1).

Although homogeneity is strongly used in our main result, one can replace it by other hypotheses that make it possible to avoid the compensation problem. The general context is then a Riemannian submersion with totally geodesic fibers,

$$F \to M \xrightarrow{\pi} B, \tag{1-4}$$

and metrics g_t on M, obtained by shrinking the fibers. Unfortunately, in general, these alternative hypotheses are quite restrictive. Since avoiding this compensation is the central issue in our results, it is natural to expect that a deeper understanding of this issue would allow weaker hypotheses; see Remark 6.2. One possibility is to impose curvature conditions that imply certain lower bounds on eigenvalues of the Laplacian, in which case bifurcation is obtained via the easier Morse index criterion.

Theorem B. *Let $F \to M \to B$ be a Riemannian submersion with totally geodesic fibers. Let $l = \dim F \geq 2$ and $m = \dim M$. Assume the metrics g_t obtained by shrinking the fibers have constant scalar curvature*[1] *and that for some $\tau > 0$ and $k_1, k_2 > 0$,*

$$\begin{cases} \mathrm{Ric}_F \geq (l-1)k_1 \\ \mathrm{scal}_F < l(m-1)k_1 \end{cases} \quad \textit{and} \quad \begin{cases} \mathrm{Ric}_{(M,g_\tau)} \geq (m-1)k_2 \\ \mathrm{scal}_B \leq m(m-1)k_2. \end{cases}$$

There exists an infinite subset $\mathcal{T}$ of positive real numbers accumulating at 0, such that for each $t \in \mathcal{T}$ there are at least 3 solutions to the Yamabe problem in the conformal class $[g_t]$.

[1] For example, this happens if the original metric on M is Einstein; see [Besse 1987, Corollary 9.62].

The paper is organized as follows. In Section 2, we briefly review the variational setup for the Yamabe problem and the bifurcation techniques (Propositions 2.2 and 2.3) introduced in [Bettiol and Piccione 2013; Lima et al. 2012]. The effect of shrinking the fibers on the spectrum of a Riemannian submersion with totally geodesic fibers is recalled in Section 3. The core of the proof of Theorem A (Theorem 4.1) is given in Section 4. Section 5 describes several examples to which these theorems apply. Section 6 contains the core of the proof of Theorem B (Theorem 6.1). Finally, Section 7 explains how to translate bifurcation of solutions into the multiplicity results claimed above.

2. Variational framework and bifurcation criteria

We start by briefly recalling the classic variational setup for the Yamabe problem; see [Bettiol and Piccione 2013; Lima et al. 2012; Schoen 1989] for details. Let M be a closed smooth manifold of dimension m. Consider the set $\mathrm{Met}(M)$ of $C^{r,\alpha}$ Riemannian metrics on M, which is an open convex cone in the Banach space of $C^{r,\alpha}$ symmetric $(0,2)$-tensors. Henceforth we fix $r \geq 3$ and $\alpha \in \,]0,1[$. For each $g \in \mathrm{Met}(M)$, define its $C^{r,\alpha}$ conformal class by

$$[g] = \{\phi\, g : \phi \in C^{r,\alpha}(M), \phi > 0\}.$$

Denote by $\mathrm{Met}_1(M)$ the smooth codimension 1 embedded submanifold of $\mathrm{Met}(M)$ formed by unit volume metrics. Finally, let

$$[g]_1 = \mathrm{Met}_1(M) \cap [g].$$

The set $[g]_1$ is a codimension 1 Banach submanifold of $[g]$, and its tangent space at the metric g can be canonically identified as

$$T_g[g]_1 \cong \left\{\psi \in C^{r,\alpha}(M) : \int_M \psi \,\mathrm{vol}_g = 0\right\}. \tag{2-1}$$

The choice of Hölder regularity $C^{r,\alpha}$ for the metric tensors and functions is due to a technical analytic aspect of our theory that employs certain *Schauder estimates*. To simplify the exposition, these technicalities will not be discussed in further detail.

Hilbert–Einstein functional. Fix a metric g_0 on the manifold M and consider the Hilbert–Einstein functional $\mathcal{A}$ defined in (1-1). The restriction of $\mathcal{A}$ to $[g_0]_1$ is smooth, and its critical points are the constant scalar curvature metrics in $[g_0]_1$. Given one such critical point $g \in [g_0]_1$, the second variation of $\mathcal{A}$ at g can be identified with the quadratic form on (2-1) given by

$$\mathrm{d}^2\big(\mathcal{A}|_{[g]_1}\big)(g)(\psi,\psi) = \frac{m-2}{2}\int_M \big((m-1)\Delta_g\psi - \mathrm{scal}(g)\psi\big)\psi \,\mathrm{vol}_g, \tag{2-2}$$

where Δ_g is the Laplace–Beltrami operator of g, or *Laplacian* of g, with the sign convention such that its spectrum is nonnegative. The above quadratic form is represented by the (formally) self-adjoint elliptic operator

$$J_g(\psi) = \frac{(m-1)(m-2)}{2}\left(\Delta_g\psi - \frac{\mathrm{scal}(g)}{m-1}\,\psi\right), \tag{2-3}$$

which we call the *Jacobi operator* at g. From the above formulas, the critical point $g \in [g_0]_1$ is *nondegenerate* (in the usual sense of Morse theory) if and only if $\mathrm{scal}(g)/(m-1)$ is not an eigenvalue of the Laplacian Δ_g, or if $\mathrm{scal}(g) = 0$. Moreover, the *Morse index* $N(g)$ of this critical point is given by the number of negative eigenvalues of (2-3), that is, the number of *positive* eigenvalues of Δ_g, counted with multiplicity, that are less than $\mathrm{scal}(g)/(m-1)$. More precisely, denote by

$$0 < \lambda_1^g \le \lambda_2^g \le \lambda_3^g \le \cdots \le \lambda_j^g \le \cdots$$

the sequence of eigenvalues of the Laplacian Δ_g, repeated according to their multiplicity. Then, the Morse index of g is given by

$$N(g) = \max\left\{ j \in \mathbb{N} : \lambda_j^g < \frac{\mathrm{scal}(g)}{m-1} \right\}. \tag{2-4}$$

Remark 2.1. For the purposes of this paper, the relevant data are the signs of the eigenvalues of the operator (2-3). Note that, given $\alpha > 0$, one has $\Delta_{\alpha g} = \frac{1}{\alpha}\Delta_g$ and $\mathrm{scal}(\alpha g) = \frac{1}{\alpha}\mathrm{scal}(g)$. Hence, the spectrum of (2-3) scales in a trivial way under homotheties, in the sense that negative (respectively positive) eigenvalues remain negative (respectively positive). On the other hand, $\mathrm{vol}_{\alpha g} = \alpha^{m/2}\mathrm{vol}_g$. Thus, whenever necessary, we may renormalize a metric to have unit volume without compromising the above spectral theory.

A classic result of variational bifurcation theory. Let us turn to the main tool used in this paper, bifurcation theory. Consider a continuous path $g_t \in \mathrm{Met}(M)$ of solutions to the Yamabe problem, as in (1-2). A value $t_* \in [a,b]$ is a *bifurcation value* for the family g_t if there exists a sequence $\{t_q\}$ in $[a,b]$ that converges to t_* and a sequence $\{g_q\}$ in $\mathrm{Met}(M)$ that converges to g_{t_*} satisfying for all $q \in \mathbb{N}$:

(i) $g_{t_q} \in [g_q]$, but $g_q \neq g_{t_q}$.

(ii) $\mathrm{Vol}(g_q) = \mathrm{Vol}(g_{t_q})$.

(iii) $\mathrm{scal}(g_q)$ is constant.

If $t_* \in [a,b]$ is *not* a bifurcation value, then the family g_t is *locally rigid* at t_*. In other words, the family g_t is locally rigid at $t_* \in [a,b]$ if there exists a neighborhood U of g_{t_*} in $\mathrm{Met}(M)$ such that, for $t \in [a,b]$ sufficiently close to t_*, the conformal class $[g_t]$ contains a unique metric of constant scalar curvature in U whose volume equals the volume of g_t.

Using a suitable version of the Implicit Function Theorem, one sees that if g_{t_*} is a nondegenerate critical point, then g_t is locally rigid at t_*; see [Lima et al. 2012, Proposition 3.1]. Thus, *degeneracy* is a necessary condition for bifurcation, however it is in general not sufficient. A classic result in variational bifurcation theory states that, given a continuous path of smooth functionals and a continuous path of critical points, there is a bifurcating branch issuing from the given path at each point where the Morse index changes (see [Kielhöfer 2004, Theorem II.7.3] or [Smoller and Wasserman 1990, Theorem 2.1]). Translating this result to our variational framework, we get the following.

Proposition 2.2 [Lima et al. 2012, Theorem 3.3]. *Let $[a,b] \ni t \mapsto g_t \in \mathrm{Met}(M)$ be a continuous path of constant scalar curvature metrics on M, and assume that a and b are not degeneracy values for g_t. If $N(g_a) \neq N(g_b)$, then there exists a bifurcation value $t_* \in \,]a,b[$ for the family g_t.*

Equivariant variational bifurcation. In many applications, the criterion of Proposition 2.2 cannot be employed because establishing a change of the Morse index at a given degeneracy value may be a difficult task. However, when the variational setup has an *equivariant* nature, one can replace the change of Morse index condition with a more general condition based on the representation theory of the group of symmetries of the variational problem. This more general condition fits perfectly the setup discussed in the present paper, and we now describe it in the variational framework of the Yamabe problem.

Suppose there exists a finite-dimensional *nice*[2] Lie group K that acts (on the left) by diffeomorphisms on a compact manifold M, and let $[a,b] \ni t \mapsto g_t \in \mathrm{Met}(M)$ be a continuous path of constant scalar curvature metrics on M. Up to a suitable normalization, let us assume that each g_t has unit volume (see Remark 2.1). The K-action on M induces a (right) K-action on $\mathrm{Met}(M)$ by pull-back; that is, the action of $k \in K$ on $g \in \mathrm{Met}(M)$ is k^*g. Assume that the K-action on M is isometric with respect to all metrics g_t, that is,

[2] A group G is *nice* in the sense of [Smoller and Wasserman 1990] if, given unitary representations π_1 and π_2 of G on Hilbert spaces V_1 and V_2 respectively, such that $B_1(V_1)/S_1(V_1)$ and $B_1(V_2)/S_1(V_2)$ have the same (equivariant) homotopy type as G-spaces, then π_1 and π_2 are unitarily equivalent. Here, B_1 and S_1 denote respectively the unit ball and the unit sphere in the specified Hilbert space, and the quotient $B_1(V_i)/S_1(V_i)$ is meant in the topological sense (that is, it denotes the unit ball of V_i with its boundary contracted to one point).

For example, any compact connected Lie group G is nice. More generally, any compact Lie group with less than 5 connected components is nice. Denoting by G^0 the identity connected component of G, then G is nice if the discrete part G/G^0 is either the product of a finite number of copies of $\mathbb{Z}_2$ (for example, the case $G = \mathrm{O}(n)$); or the product of a finite number of copies of $\mathbb{Z}_3$; or if $G/G^0 = \mathbb{Z}_4$.

(2-5) $$k^*(g_t) = g_t \quad \text{for all } k \in K,\ t \in [a,b].$$

In this situation, the K-action on $\mathrm{Met}(M)$ leaves every conformal class $[g_t]$ invariant, and also $[g_t]_1$ is K-invariant for all t. Note that (2-5) means that, for all t, g_t is a fixed point of the K-action on $\mathrm{Met}(M)$.

It is easy to see that, given $\phi\colon M \to \mathbb{R}$ a positive function,

$$\mathcal{A}\big(\phi^{4/(n-2)}\, g_t\big) = \int_M \left(4\,\frac{n-1}{n-2}\,\phi\,\Delta_t\phi + \mathrm{scal}(g_t)\phi^2\right)\mathrm{vol}_{g_t},$$

where Δ_t is the Laplacian of g_t. Using that $k^*(\phi\, g_t) = (\phi\circ k)g_t$, right-composition with isometries commutes with Δ_t and $\mathrm{scal}(g_t)$ is constant, it follows from a change of variables that

$$\mathcal{A}\big(k^*(\phi\, g_t)\big) = \mathcal{A}\big(\phi\, g_t\big) \quad \text{for all } k \in K,\ t \in [a,b].$$

Thus, denoting by $\mathcal{A}_t$ the restriction of the Hilbert–Einstein functional (1-1) to the conformal class $[g_t]_1$, we have that $\mathcal{A}_t$ is invariant under the K-action on $[g_t]_1$.

For each eigenvalue $\lambda \in \mathrm{Spec}(\Delta_t)$, denote by $E_t^\lambda \subset L^2(M, \mathrm{vol}_{g_t})$ the correspondent eigenspace. Elements of E_t^λ are smooth functions, and $\dim E_t^\lambda < +\infty$ is the multiplicity of λ. It is easy to see that, for every $t \in [a,b]$ and $\lambda \in \mathrm{Spec}(\Delta_t)$, there is a representation π_t^λ of K in E_t^λ, given by right-composition with isometries:

(2-6) $$\pi_t^\lambda(k)\psi = \psi\circ k \quad \text{for all } k \in K,\ \psi \in E_t^\lambda.$$

Note this is (the restriction to E_t^λ of) the isotropy representation of the K-action on $[g_t]_1$ at the fixed point g_t, that is, the linearization of this K-action at g_t. We remark that since the K-action by pull-back on $[g_t]_1$ is a *right* action, its linearization at a fixed point is actually an *anti*representation. However, we will henceforth not make distinctions between left/right actions and representations/antirepresentations since this does not affect our arguments.

Denote by $\mathcal{N}_t$ the set of negative eigenvalues of the operator (2-3); that is,

$$\mathcal{N}_t := \mathrm{Spec}\big(J_{g_t}\big) \cap\,]-\infty, 0[.$$

Now, for all $t \in [a,b]$, define the *negative isotropic representation* π_t^- of K in $\bigoplus_{\lambda\in\mathcal{N}_t} E_t^\lambda$ as the direct sum representation

$$\pi_t^- := \bigoplus_{\lambda\in\mathcal{N}_t} \pi_t^\lambda.$$

Notice that the degree[3] of π_t^- is always finite and equal to the Morse index $N(g_t)$ defined in (2-4). Finally, recall that two representations $\pi_i\colon K \to \mathrm{GL}(V_i)$, $i = 1, 2$,

[3]That is, the dimension of the vector space $\bigoplus_{\lambda\in\mathcal{N}_t} E_t^\lambda$.

are *isomorphic* if there exists a K-equivariant isomorphism $T\colon V_1 \to V_2$, that is, such that $\pi_2(k) \circ T = T \circ \pi_1(k)$ for all $k \in K$.

Proposition 2.3. *Let K be a nice Lie group acting on a compact manifold M and $[a,b] \ni t \mapsto g_t \in \mathrm{Met}(M)$ be a continuous path of (unit volume) constant scalar curvature metrics. Suppose the K-action is isometric on (M, g_t) for all $t \in [a,b]$. If a and b are not degeneracy values for g_t, and if the negative isotropic representations π_a^- and π_b^- are not isomorphic, then there exists a bifurcation value $t_* \in \,]a,b[$ for the family g_t.*

Proof. This result is a direct application of an equivariant bifurcation result due to Smoller and Wasserman [1990, Theorem 3.1]. More precisely, one needs a slightly more general statement of the result, applied to functionals defined on a varying manifold. In our case, the functional is the Hilbert–Einstein functional $\mathcal{A}$, defined on the varying manifold $[g_t]_1$. Details can be found in [Lima et al. 2012, Theorem 3.4, Theorem A.2]. □

Notice the above is a refinement of Proposition 2.2, which corresponds to saying that π_a^- and π_b^- do not have the same degree (and hence cannot be isomorphic).

3. Laplacian on collapsing Riemannian submersions

In order to study bifurcation from an initial family g_t of solutions to the Yamabe problem on M, it is crucial to have a good understanding of the spectra of their Laplacians Δ_t. In all our applications, the family g_t is obtained as a deformation of the original metric g on the total space M of a Riemannian submersion with totally geodesic fibers (1-4), by multiplying it by a factor t^2 in the direction of the fibers. We will be particularly interested in the behavior of the spectrum of Δ_t as the fibers collapse to a point, that is, as $t \to 0$.

The effect of such deformation on the spectrum was first studied by Bérard-Bergery and Bourguignon [1982], where g_t is called *canonical variation* of g. The starting point is that (1-4) remains a Riemannian submersion with totally geodesic fibers when g is replaced by g_t; see [Bérard-Bergery and Bourguignon 1982, Proposition 5.2]. For the readers' convenience, we briefly recall some related results that are discussed in more detail in [Bettiol and Piccione 2013, Section 3].

Vertical Laplacian and lifts of eigenfunctions. Define the *vertical Laplacian* Δ_v on a function $\psi\colon M \to \mathbb{R}$ by

$$(\Delta_v \psi)(p) := (\Delta_F \psi|_{F_p})(p) \quad \text{for all } p \in M,$$

where Δ_F is the Laplacian of the fiber and $F_p = \pi^{-1}(\pi(p))$ is the fiber through $p \in M$. Just like a usual Laplacian, the vertical Laplacian is a nonnegative self-adjoint unbounded operator on $L^2(M)$, however it is *not elliptic* (unless π is a covering). Since the fibers are isometric, Δ_v has a discrete spectrum equal to that of the Laplacian Δ_F of the fiber. Let us denote

$$\begin{aligned}\operatorname{Spec}(\Delta_M) &= \{0=\mu_0<\mu_1<\cdots<\mu_k \nearrow +\infty\},\\ \operatorname{Spec}(\Delta_v) &= \{0=\phi_0<\phi_1<\cdots<\phi_j \nearrow +\infty\},\end{aligned} \tag{3-1}$$

where these eigenvalues are *not repeated* according to their multiplicity. Note that the multiplicity of the eigenvalues of Δ_M is always *finite*, but the eigenvalues of Δ_v might have *infinite* multiplicity. For instance, $\Delta_v\widetilde{\psi}=0$ implies only that $\widetilde{\psi}$ is constant *along the fibers*; that is, $\widetilde{\psi}=\psi\circ\pi$ for some function $\psi\colon B\to\mathbb{R}$ on the base.

It is easy to see that, for any $\psi\colon B\to\mathbb{R}$ and its *lift* $\widetilde{\psi}:=\psi\circ\pi$,

$$\Delta_M\widetilde{\psi}=(\Delta_B\psi)\circ\pi+g\big(\operatorname{grad}_g\widetilde{\psi},\vec{H}\big), \tag{3-2}$$

where $\vec{H}$ is the mean curvature vector field of the fibers. Since we assumed the fibers of π are totally geodesic, $\vec{H}$ vanishes identically. Thus, if ψ is an eigenfunction of Δ_B, then its lift $\widetilde{\psi}$ is an eigenfunction of Δ_M with the same eigenvalue (and constant along the fibers). Therefore, there is a natural inclusion

$$\operatorname{Spec}(\Delta_B)\subset\operatorname{Spec}(\Delta_M). \tag{3-3}$$

Conversely, if $\psi\colon M\to\mathbb{R}$ is constant along the fibers and satisfies $\Delta_M\psi=\lambda\psi$, then there exists $\check{\psi}\colon B\to\mathbb{R}$ such that $\Delta_B\check{\psi}=\lambda\check{\psi}$ and $\psi=\check{\psi}\circ\pi$. Summing up, it follows from (3-2), after checking the adequate regularity hypotheses (see [Bessa et al. 2012, Lemma 3.11]), that the following holds.

Proposition 3.1. *If $\pi: M\to B$ is a Riemannian submersion with totally geodesic fibers, then an eigenfunction of Δ_M is constant along the fibers if and only if it is the lift of an eigenfunction of Δ_B.*

Spectrum of deformed metrics. Another consequence of having totally geodesic fibers is that $L^2(M)$ has a Hilbert basis consisting of simultaneous eigenfunctions of the original Laplacian Δ_M and the vertical Laplacian Δ_v; see [Bérard-Bergery and Bourguignon 1982, Theorem 3.6]. This means that these operators can be *simultaneously diagonalized*, in the appropriate sense. Furthermore, it is a simple calculation to show that that the Laplacian Δ_t of the deformed metric g_t is $\Delta_t=\Delta_M+(1/t^2-1)\Delta_v$; see [Bérard-Bergery and Bourguignon 1982, Proposition 5.3]. From this, we get the following description of its spectrum.

Proposition 3.2 [Bettiol and Piccione 2013, Corollary 3.6]. *For each $t>0$, the following inclusion holds*:

$$\operatorname{Spec}(\Delta_t)\subset\operatorname{Spec}(\Delta_M)+(1/t^2-1)\operatorname{Spec}(\Delta_v).$$

Since these sets are discrete, every eigenvalue $\lambda(t)$ *of* Δ_t *is of the form*

$$\lambda^{k,j}(t) = \mu_k + (1/t^2 - 1)\phi_j, \tag{3-4}$$

for some eigenvalues μ_k *and* ϕ_j *of* Δ_M *and* Δ_v, *respectively.*

Remark 3.3. Not all possible combinations of μ_k and ϕ_j in (3-4) give rise to an eigenvalue of Δ_t. In fact, this only happens when the total space of the submersion is a Riemannian product. In general, determining which combinations are allowed depends on the global geometry of the submersion.

Note that, since the fibers of π remain totally geodesic with respect to g_t, (3-3) holds when Δ_M is replaced with Δ_t; that is,

$$\mathrm{Spec}(\Delta_B) \subset \mathrm{Spec}(\Delta_t) \quad \text{for all } t > 0. \tag{3-5}$$

Moreover, when $j = 0$ in (3-4), $\lambda^{k,0}(t) = \mu_k \in \mathrm{Spec}(\Delta_M)$ remains an eigenvalue of Δ_t for $t \neq 1$ if and only if $\mu_k \in \mathrm{Spec}(\Delta_B)$. Such eigenvalues $\lambda^{k,0}(t)$ of Δ_t will be called *constant eigenvalues*, since they are independent of t. In other words, the constant eigenvalues of Δ_t are the ones in the left-hand side of (3-5). We stress that $\lambda^{k,0}(t)$ is not necessarily a constant eigenvalue for all k.

4. Bifurcation on homogeneous fibrations

Let $H \subsetneq K \subsetneq G$ be compact connected Lie groups such that $\dim K/H \geq 2$, and assume that either H is normal in K, or K is normal in G. Consider the homogeneous fibration (1-3),

$$K/H \to G/H \xrightarrow{\pi} G/K, \quad \text{where } \pi(gH) = gK,$$

and notice that the fiber over $gK \in G/K$ is $(gK)H \subset G/H$. Define a K-action on G/H by $k \cdot gH = kgH$ if K is normal in G and by $k \cdot gH = gk^{-1}H$ if H is normal in K. Notice that the orbits of this K-action are exactly the fibers of π. Denote by $\mathfrak{h} \subsetneq \mathfrak{k} \subsetneq \mathfrak{g}$ the Lie algebras of $H \subsetneq K \subsetneq G$. We henceforth fix an $\mathrm{Ad}_G(K)$-invariant complement $\mathfrak{m}$ to $\mathfrak{k}$ in $\mathfrak{g}$, and an $\mathrm{Ad}_G(H)$-invariant complement $\mathfrak{p}$ to $\mathfrak{h}$ in $\mathfrak{k}$; that is,

$$\mathfrak{k} \oplus \mathfrak{m} = \mathfrak{g}, \quad [\mathfrak{k}, \mathfrak{m}] \subset \mathfrak{m} \quad \text{and} \quad \mathfrak{h} \oplus \mathfrak{p} = \mathfrak{k}, \quad [\mathfrak{h}, \mathfrak{p}] \subset \mathfrak{p}.$$

There are natural identifications of $\mathfrak{m}$ and $\mathfrak{p}$ with the tangent spaces to G/K and K/H at the origin,[4] respectively. The sum $\mathfrak{m} \oplus \mathfrak{p}$ is an $\mathrm{Ad}_G(H)$-invariant complement to $\mathfrak{h}$ in $\mathfrak{g}$, which is identified with the tangent space to G/H at the origin.

Any $\mathrm{Ad}_G(K)$-invariant inner product on $\mathfrak{m}$ defines a G-invariant metric on the base G/K, and any $\mathrm{Ad}_G(H)$-invariant inner product on $\mathfrak{p}$ defines a K-invariant

[4] That is, $\mathfrak{m} \cong T_{(eK)}G/K$ and $\mathfrak{p} \cong T_{(eH)}K/H$.

metric on the fiber K/H. The orthogonal direct sum of these inner products on $\mathfrak{m}\oplus\mathfrak{p}$ now gives a G-invariant metric on G/H. We call these metrics on (1-3) *compatible* homogeneous metrics. We stress that not necessarily all G-invariant metrics on G/H arise in this way. However, if $\mathfrak{m}$ and $\mathfrak{p}$ do not share equivalent $\mathrm{Ad}_G(H)$-submodules, then all G-invariant metrics on G/H that project to a G-invariant metric on G/K are of this form. We will henceforth consider all homogeneous fibrations to be endowed with compatible metrics. The homogeneous fibration $\pi\colon G/H\to G/K$ is then automatically a Riemannian submersion with totally geodesic fibers (isometric to K/H); see [Bérard-Bergery 1978, Proposition 2] or [Besse 1987, Theorem 9.80].

Theorem 4.1. *Let $K/H\to G/H\to G/K$ be a homogeneous fibration as above, and assume that K/H has positive scalar curvature. Let g_t be the family of homogeneous metrics on G/H obtained by scaling the fibers by t^2. There exists a sequence $\{t_q\}$ in $]0,1[$, that converges to 0, of bifurcation values for the family g_t.*

Proof. Since $\pi\colon(G/H,g_t)\to G/K$ is a Riemannian submersion with totally geodesic fibers, we have

$$\mathrm{scal}(G/H,g_t)=\frac{1}{t^2}\mathrm{scal}(K/H)+\mathrm{scal}(G/K)\circ\pi-t^2\|A\|^2, \tag{4-1}$$

where $\|A\|$ is the Hilbert–Schmidt norm of the Gray–O'Neill tensor of integrability of the horizontal distribution (see [Besse 1987, Proposition 9.70]). Note that $\mathrm{scal}(K/H)>0$ implies

$$\lim_{t\to 0_+}\mathrm{scal}(G/H,g_t)=+\infty. \tag{4-2}$$

Recall that the *degeneracy values* in this setup are those $t>0$ such that

$$(m-1)\Delta_t\psi-\mathrm{scal}(G/H,g_t)\psi=0 \tag{4-3}$$

has a nontrivial solution ψ, where $m=\dim G/H$ and Δ_t is the Laplacian of $(G/H,g_t)$. From (3-4) and (4-1), the set of such degeneracy values is discrete. From (4-2), there are infinitely many degeneracy values t_q accumulating at 0 such that $\mathrm{scal}(G/H,g_{t_q})/(m-1)\in\mathrm{Spec}(\Delta_{G/K})\subset\mathrm{Spec}(\Delta_t)$; see (3-5). We claim that every such value t_q is a bifurcation value.

Fix one such t_q, and denote by $\lambda\in\mathrm{Spec}(\Delta_{G/K})$ the constant eigenvalue of Δ_t such that $\mathrm{scal}(G/H,g_{t_q})/(m-1)=\lambda$. If there is a change in the Morse index at t_q, that is, for $\varepsilon>0$ sufficiently small, $N(g_{t_q-\varepsilon})\neq N(g_{t_q+\varepsilon})$, then by Proposition 2.2, t_q is a bifurcation value. However, if the Morse index does not change, there must be a *compensation* of eigenvalues. Namely, there must exist nonconstant eigenvalues $\lambda^{k_1,j_1}(t),\dots,\lambda^{k_n,j_n}(t)$ of Δ_t, whose combined multiplicity equals the multiplicity of λ, such that

$$\lambda < \mathrm{scal}(G/H, g_t)/(m-1) < \lambda^{k_i, j_i}(t) \quad \text{for all } t < t_q \text{ (close to } t_q\text{) and } 1 \le i \le n,$$
$$\lambda > \mathrm{scal}(G/H, g_t)/(m-1) > \lambda^{k_i, j_i}(t) \quad \text{for all } t > t_q \text{ (close to } t_q\text{) and } 1 \le i \le n.$$

Denoting by E_t^{α} the eigenspace of the eigenvalue $\alpha \in \mathrm{Spec}(\Delta_t)$, we have the negative isotropic representations π_t^- on the linear spaces (of same finite dimension):

$$\begin{aligned} & E \oplus E_t^{\lambda} && \text{for } t < t_q \text{ (close to } t_q\text{)}, \\ & E \oplus \bigoplus_i E_t^{\lambda^{k_i, j_i}} && \text{for } t > t_q \text{ (close to } t_q\text{)}, \end{aligned} \tag{4-4}$$

where E is the space spanned by the eigenfunctions with eigenvalues less than $\mathrm{scal}(G/H, g_t)/(m-1)$ for t close to t_q. We claim that for small $\varepsilon > 0$, the negative isotropic representations $\pi_{t_q-\varepsilon}^-$ and $\pi_{t_q+\varepsilon}^-$ on the spaces (4-4) *cannot be isomorphic*. From Proposition 2.3, it then follows that t_q is a bifurcation value, concluding the proof.

Let us verify the above claim. Given a representation π of a compact group, denote by $\mathfrak{I}(\pi)$ the number of copies of the trivial representation in the irreducible decomposition of π. It is easily seen that a necessary condition for the two representations π_a and π_b to be isomorphic is that $\mathfrak{I}(\pi_a) = \mathfrak{I}(\pi_b)$. For the negative isotropic representation π_t^-, one can compute

$$\mathfrak{I}(\pi_t^-) = \sum_{\substack{\eta \in \mathrm{Spec}(\Delta_{G/K}) \\ \eta < \mathrm{scal}(g_t)/(m-1)}} \mathrm{mul}(\eta), \tag{4-5}$$

where $\mathrm{mul}(\eta)$ is the multiplicity of η as an eigenvalue of $\Delta_{G/K}$. Indeed, an eigenfunction ψ of Δ_t is constant along the fibers K/H of the homogeneous fibration (1-3) if and only if it is K-invariant, that is, $\psi \circ k = \psi$ for all $k \in K$. This is equivalent to saying that ψ is a fixed point of π_t^-; see (2-6). So, from Proposition 3.1, the left-hand side of (4-5) is greater than or equal to the right-hand side. Conversely, it is easy to see that if $\psi : G/H \to \mathbb{R}$ is the linear combination of eigenfunctions $\psi_i : G/H \to \mathbb{R}$ of Δ_t, and if ψ is constant along the fibers of $G/H \to G/K$, then each ψ_i must be constant along such fibers. This follows from the fact that the subspace of $L^2(G/H)$ of functions that are constant along the fibers (which is isomorphic to $L^2(G/K)$) is spanned by the (lift of) eigenfunctions of $\Delta_{G/K}$. In other words, the space spanned by the eigenfunctions of $\Delta_{G/H}$ that are constant along the fibers and the space spanned by the eigenfunctions of $\Delta_{G/H}$ that are not constant along the fibers are L^2-orthogonal. This completes the proof of (4-5).

From (4-4) and (4-5) we see that, for any $\varepsilon > 0$ small, $\mathfrak{I}(\pi_{t_q-\varepsilon}^-) > \mathfrak{I}(\pi_{t_q+\varepsilon}^-)$. Therefore these representations are not isomorphic, concluding the proof. □

Remark 4.2. At all bifurcation values for the family g_t a *break of symmetry* occurs, in the sense that any solutions in the bifurcating branch are not G-homogeneous. This follows easily from the fact that each conformal class contains at most one homogeneous metric (up to rescaling).

Sharpness of fiber hypotheses. If the fibers K/H have flat scalar curvature or have dimension 1, then $\mathrm{scal}(g_t)$ remains bounded as $t \to 0$, and there are not infinitely many degeneracy values as above. For instance, consider a fibration of tori, $G = K \times K$, $K = T^2$ and $H = \{e\}$, where the inclusion of K is as one of the factors of G. If K is endowed with the flat metric and G with the product metric, shrinking the fibers keeps the total space G/H flat, hence the family obtained is (trivially) locally rigid, for all $t > 0$.

5. Examples

We now discuss how to construct examples of homogeneous fibrations with fibers of positive scalar curvature to which Theorem 4.1 (hence also Theorem A) applies.

Normal homogeneous metrics. A K-invariant metric on K/H is called *normal* if it is obtained from the restriction to $\mathfrak{p}$ of a bi-invariant inner product on $\mathfrak{k}$. Since K is compact, it admits a bi-invariant metric. Hence, normal homogeneous metrics always exist[5] on K/H. Endowed with such a metric, the sectional curvature of a tangent plane at the origin, spanned by orthonormal vectors X and Y, is

$$\mathrm{sec}(X, Y) = \tfrac{1}{4}\left\|[\overline{X}, \overline{Y}]\right\|^2 + \tfrac{3}{4}\left\|[\overline{X}, \overline{Y}]_{\mathfrak{h}}\right\|^2 \geq 0, \tag{5-1}$$

where $\overline{X} = (0, X)$ and $\overline{Y} = (0, Y)$ are the horizontal lifts of X and Y to $\mathfrak{k} = \mathfrak{h} \oplus \mathfrak{p}$ and $(\,\cdot\,)_{\mathfrak{h}}$ denotes the $\mathfrak{h}$-component of a vector in $\mathfrak{k}$. In particular, $\mathrm{scal}(K/H) \geq 0$, and it is equal to zero if and only if $\mathfrak{p}$ is an abelian ideal of $\mathfrak{k}$. This, in turn, is equivalent to the existence of an abelian subgroup $A \subset K$ that acts transitively on K/H. Since (the closure of) A is a compact connected abelian Lie group, it must be a torus. Hence, K/H itself is a torus. Thus, *any* normal metric on K/H has positive scalar curvature, unless K/H is a torus.

More generally, K/H admits a (normal) metric of positive scalar curvature if and only if its universal covering is not diffeomorphic to an Euclidean space; see [Bérard-Bergery 1978, Theorem 2]. In particular, if K/H is not an aspherical manifold, then K/H has a metric of positive scalar curvature. We also stress that, generally, there are other K-invariant metrics (not necessarily normal) on K/H that have positive scalar curvature, and any such metric can be used to obtain

[5] There exist homogeneous spaces on which *all* homogeneous metrics are normal. These spaces are called of *normal type* in [Bérard-Bergery 1978, Section 7]. Examples are spaces whose isotropy representation is irreducible (for example, irreducible symmetric spaces) and their products.

examples of applications of our results. In this direction, examples with nonnormal homogeneous metrics with positive scalar curvature on spheres will be discussed below.

In any of the cases above, one can endow the remaining spaces of (1-3) with compatible homogeneous metrics. In this way, it is possible to construct many classes of homogeneous fibrations to which our results apply.

Twisted products. Let us now describe explicit triples $H \subsetneq K \subsetneq G$ of compact connected Lie groups with either H normal in K or K normal in G. Starting with the latter, if a compact connected Lie group G has a proper connected normal subgroup K, then there exists another connected normal subgroup[6] L of G such that $G = (K \times L)/\Gamma$, where $\Gamma \subset K \times L$ is finite. For any subgroup H of K, one then gets the homogeneous fibration

$$K/H \to ((K \times L)/\Gamma)/H \to G/K. \tag{5-2}$$

This provides an algorithm to build examples, whose input are the groups H, K, L and Γ. Setting $G = (K \times L)/\Gamma$ we then have that the factor K is a normal subgroup.

Example 5.1. Consider $G = \mathrm{SO}(4)$, which is double-covered by $S^3 \times S^3$. In this case, $K = L = S^3$ and $\Gamma = \mathbb{Z}_2$ is the diagonal embedding into $K \times L$; that is, Γ is the subgroup generated by $(-1, -1) \in S^3 \times S^3$. We then have

$$G = (K \times L)/\Gamma = (S^3 \times S^3)/\mathbb{Z}_2,$$

and the quotient G/K is isomorphic to $S^3/\mathbb{Z}_2 = \mathrm{SO}(3)$. One can choose $H \subset K = S^3$ to be trivial, so that $K/H = S^3$; or to be a circle, for example the circle that gives the Hopf action on S^3, so that $K/H = S^2$. The corresponding homogeneous fibrations that (5-2) gives are

$$S^3 \to \mathrm{SO}(4) \to \mathrm{SO}(3) \quad \text{and} \quad S^2 \to \mathrm{SO}(4)/S^1 \to \mathrm{SO}(3).$$

In these cases the total space G/H is a *twisted product*, while in the special case where Γ is trivial we have the splitting $G = K \times L$, so that $G/K = L$ and $G/H = K/H \times L$ is an actual *product* manifold. That is, (5-2) becomes

$$K/H \to K/H \times L \to L. \tag{5-3}$$

[6]The subgroup L is obtained as the connected subgroup of G whose Lie algebra $\mathfrak{l}$ is the orthogonal complement (with respect to a bi-invariant metric) of the Lie algebra $\mathfrak{k}$ of K. Since K is normal, $\mathfrak{k}$ is an ideal and then $\mathfrak{l} = \mathfrak{k}^{\perp}$ is also an ideal, because the adjoint representation is skew-symmetric with respect to the bi-invariant metric. This implies that L is normal, and KL generates the entire group G by connectedness, since it does so locally near the identity. Finiteness of $\Gamma = K \cap L$ follows since the intersection $\mathfrak{k} \cap \mathfrak{l}$ is trivial.

In contrast to the cases above, the deformed metrics g_t in this situation are product metrics, obtained by rescaling the directions tangent to the first factor K/H by t^2 and keeping the metric constant in the directions tangent to L. In this way, any product of a compact homogeneous space (with positive scalar curvature) and a compact connected Lie group satisfies the hypotheses of Theorem 4.1 (hence also of Theorem A). We note that this particular case is covered by the results of [Lima et al. 2012].

Sphere fibers. An important observation is that in the above product situation where $K/H = S^n$ is a sphere, our result allows for *any* homogeneous metric (not necessarily normal) on $K/H = S^n$ whose scalar curvature is positive, as opposed to only the round metric. Homogeneous metrics on spheres were classified by Ziller [1982]; they are obtained by rescaling the fibers of one of the Hopf fibrations:

$$S^1 \to S^{2n+1} \to \mathbb{C}P^n, \quad S^3 \to S^{4n+3} \to \mathbb{H}P^n, \quad S^7 \to S^{15} \to S^8\big(\tfrac{1}{2}\big).$$

In the first and last case, there is only one direction in which the fibers can be rescaled, while for the fibration with S^3 fiber, each of the 3 coordinate S^1 subgroups can be rescaled with a different factor. This gives rise to the following metrics:

- $\mathtt{g}_s$, a 1-parameter family of $\mathrm{U}(n+1)$-invariant metrics on S^{2n+1};
- $\mathtt{h}_{s_1,s_2,s_3}$, a 3-parameter family of $\mathrm{Sp}(n+1)$-invariant metrics on S^{4n+3};
- $\mathtt{k}_s$, a 1-parameter family of $\mathrm{Spin}(9)$-invariant metrics on S^{15}.

All these metrics have $\mathrm{scal} > 0$ for a certain range of parameters (see table below). Thus, if $K/H = S^n$ is isometric to one of those spheres, the submersion (5-3) satisfies the hypotheses of Theorem 4.1 (hence also of Theorem A).

metric	K	H	$\dim K/H$
round	$\mathrm{SO}(n+1)$	$\mathrm{SO}(n)$	$n\ (\geq 2)$
	$\mathrm{Spin}(7)$	G_2	7
	G_2	$\mathrm{SU}(3)$	6
$\mathtt{g}_s$	$\mathrm{SU}(n+1)$	$\mathrm{SU}(n)$	$2n+1\ (\geq 2)$
	$\mathrm{U}(n+1)$	$\mathrm{U}(n)$	$2n+1\ (\geq 2)$
$\mathtt{h}_{s_1,s_2,s_3}$	$\mathrm{Sp}(n+1)$	$\mathrm{Sp}(n)$	$4n+3$
	$\mathrm{Sp}(n+1)\times\mathrm{Sp}(1)$	$\mathrm{Sp}(n)\times\mathrm{Sp}(1)$	$4n+3$
	$\mathrm{Sp}(n+1)\times\mathrm{U}(1)$	$\mathrm{Sp}(n)\times\mathrm{U}(1)$	$4n+3$
$\mathtt{k}_s$	$\mathrm{Spin}(9)$	$\mathrm{Spin}(7)$	15

It is proved in [Bettiol and Piccione 2013, Proposition 4.2] that the range of parameters for which these K-invariant metrics have $\mathrm{scal} > 0$ is $0 < s < s_{\max}$, where

$$s_{\max} = \begin{cases} \sqrt{2n+2} & \text{in the case of } \mathrm{g}_s, \\ \sqrt{\frac{2n+4}{3} + \frac{\sqrt{18n+16(n^2+2n)^2}}{6n}} & \text{in the case of } \mathtt{h}_{s,s,s}, \\ \sqrt{2+\frac{1}{2}\sqrt{19}} & \text{in the case of } \mathtt{k}_s. \end{cases}$$

When $s = 1$, these metrics are isometric to the round metric (which is the only normal homogeneous metric in each family).

Remark 5.2. The above construction can be interpreted as having a chain

$$H \subsetneq H' \subsetneq K \subsetneq G$$

of Lie groups, and first performing a Cheeger deformation with respect to the H'-action in the total space of $H'/H \to K/H \to K/H'$ to obtain positive scalar curvature on K/H, and then using this metric on the fiber of $K/H \to G/H \to G/K$. The Cheeger deformation with respect to the K-action on G/H gives the desired 1-parameter family g_t that satisfies the hypotheses of our results. More generally, one could perform multiple "preliminary" Cheeger deformations with a longer chain of groups in order to gain $\mathrm{scal} > 0$ on the fibers of the "last" homogeneous fibration.

Example 5.3. Another interesting class of examples with K normal in G is when H is trivial, so that the resulting homogeneous fibration (1-3) is a short exact sequence of Lie groups $K \to G \to G/K$. This is precisely the case of $S^3 \to \mathrm{SO}(4) \to \mathrm{SO}(3)$ in Example 5.1. Here, the deformed metrics g_t are obtained by shrinking the original metric in the direction of the cosets of K in G.

Other examples. As explained above, instead of having K normal in G, one can also consider the case where H is normal in K. This poses far fewer restrictions on the homogeneous fibrations that can be obtained, since the group K will split as a product (up to a finite quotient); however, no conditions are imposed on G. For instance, G may have arbitrarily large dimension and rank. It follows from the above discussion on normal homogenous metrics that our results apply to the submersion (1-3) with H normal in K as soon as the quotient K/H is not abelian. More precisely, if K/H is not a torus, then any normal homogeneous metric will have positive scalar curvature. Any choice of G will then yield a triple $H \subsetneq K \subsetneq G$ whose corresponding homogeneous fibration can be endowed with compatible metrics for which Theorem 4.1 (hence also Theorem A) applies.

Example 5.4. To illustrate the above comments, let us build on the case described in Example 5.1. Instead of G, set $K = \mathrm{SO}(4) = (S^3 \times S^3)/\mathbb{Z}_2$ and H as one of the S^3

factors, so that $K/H = \mathrm{SO}(3)$. Then G can be chosen arbitrarily among compact connected Lie groups that have a subgroup isomorphic to $\mathrm{SO}(4)$. As concrete examples, we may set $G = \mathrm{SO}(5)$ so that $G/K = S^4$ is a sphere; or $G = \mathrm{SO}(6)$ so that $G/K = T_1S^4$ is the unit tangent bundle of S^4. The corresponding homogeneous fibrations are

$$\mathrm{SO}(3) \to \mathrm{SO}(5)/S^3 \to S^4 \quad \text{and} \quad \mathrm{SO}(3) \to \mathrm{SO}(6)/S^3 \to T_1S^4.$$

Remark 5.5. Any of the above examples can be trivially used to obtain new ones with nonsimply connected total space. Consider $F \subset K$ a finite subgroup and its action on K/H and G/H, so that the inclusion map $K/H \to G/H$ is equivariant. One can form a new fibration replacing K/H and G/H by their (nonsimply connected) quotients by the F-action. Since $F \subset K$, the base of the fibration remains G/K. If the original metrics satisfied the conditions of Theorem 4.1, then the induced metrics in the new fibration also do.

6. Bifurcation on nonhomogeneous fibrations

A natural question is how the presence of many symmetries affects the bifurcation result obtained above. Homogeneity played a pivotal role in employing the equivariant bifurcation criterion (Proposition 2.3). When this assumption is dropped, the only tool at hand is the Morse index criterion (Proposition 2.2), so extra hypotheses are needed to guarantee a change in the Morse index at the degeneracy values. One such possibility is to impose certain curvature conditions that allow us to bound (from below) the growth of the eigenvalues of a nonhomogeneous collapsing Riemannian submersion.

Theorem 6.1. *Let $F \to M \to B$ be a Riemannian submersion with totally geodesic fibers. Let $l = \dim F \geq 2$ and $m = \dim M$. Assume the metrics g_t obtained by shrinking the fibers have constant scalar curvature, and that for some $\tau > 0$ and $k_1, k_2 > 0$,*

$$\begin{cases} \mathrm{Ric}_F \geq (l-1)k_1 \\ \mathrm{scal}_F < l(m-1)k_1 \end{cases} \quad \textit{and} \quad \begin{cases} \mathrm{Ric}_{(M,g_\tau)} \geq (m-1)k_2 \\ \mathrm{scal}_B \leq m(m-1)k_2. \end{cases}$$

There exists a sequence $\{t_q\}$ in $]0, \tau[$, that converges to 0, of bifurcation values for the family g_t.

Proof. Since $\pi : M \to F$ is a Riemannian submersion with totally geodesic fibers,

$$\mathrm{scal}(M, g_t) = \frac{1}{t^2}\mathrm{scal}_F + \mathrm{scal}_B \circ \pi - t^2\|A\|^2, \tag{6-1}$$

where $\|A\|$ is the Hilbert–Schmidt norm of the Gray–O'Neill tensor A. From $\mathrm{scal}_F > 0$, we have $\lim_{t\to 0_+} \mathrm{scal}(M, g_t) = +\infty$. As before, the set of degeneracy values is discrete, and infinitely many of them occur due to $\mathrm{Spec}(\Delta_B) \subset \mathrm{Spec}(\Delta_t)$;

see (3-5). Denote by t_q the sequence of degeneracy values, accumulating at 0, such that $\mathrm{scal}(M, g_{t_q})/(m-1) \in \mathrm{Spec}(\Delta_B)$. We claim that for q sufficiently large (that is, t_q sufficiently small), t_q is a bifurcation value.

From Proposition 2.2, we must verify that, for t_q sufficiently small, there is a change of the Morse index $N(g_t)$ at t_q. It suffices to prove that every *nonconstant* eigenvalue $\lambda^{k,j}(t)$ of Δ_t is strictly larger than $\mathrm{scal}(M, g_t)/(m-1)$ for t sufficiently small, so that no compensation of eigenvalues can occur (compare the proof of Theorem 4.1). Up to a simple rescaling, assume $\tau = 1$. Since the eigenvalues μ_k of Δ_M and ϕ_j of Δ_v are ordered to be monotonically increasing, it suffices to prove

$$\mathrm{scal}(M, g_t)/(m-1) < \lambda^{1,1}(t) = \mu_1 + \left(\frac{1}{t^2} - 1\right)\phi_1 \quad \text{for } t \text{ sufficiently small;} \tag{6-2}$$

see Proposition 3.2. From the Lichnerowicz estimates, since $\mathrm{Ric}_F \geq (l-1)\,k_1$ and $\mathrm{Ric}_M \geq (m-1)\,k_2$, we have

$$\phi_1 \geq l\,k_1 \quad \text{and} \quad \mu_1 \geq m\,k_2; \tag{6-3}$$

see [Chavel 1984, Chapter 3, Theorem 9]. When we combine the latter with $\mathrm{scal}_B \leq m(m-1)\,k_2$, we get

$$\mathrm{scal}_B \circ \pi - t^2\|A\|^2 \leq \mathrm{scal}_B \circ \pi \leq (m-1)\mu_1. \tag{6-4}$$

Also, from (6-3) and $\mathrm{scal}_F < l(m-1)\,k_1$, it follows that $\mathrm{scal}_F < (m-1)\phi_1$. Thus, for t sufficiently small, we have $\mathrm{scal}_F < (1-t^2)(m-1)\phi_1$, hence

$$\frac{1}{t^2}\mathrm{scal}_F < (m-1)\left(\frac{1}{t^2} - 1\right)\phi_1. \tag{6-5}$$

Adding (6-4) and (6-5) and using (6-1), we obtain (6-2), concluding the proof. □

The above result can be applied, for example, to low-dimensional Hopf fibrations, reobtaining the conclusion of Theorem 4.1. Nevertheless, for larger dimensions, the curvature pinching conditions are not satisfied, although the result remains true.

Remark 6.2. The curvature pinching conditions above are solely needed to avoid compensation of eigenvalues, in a rather forceful way. Given a Riemannian submersion $F \to M \to B$ with totally geodesic fibers of positive scalar curvature, suppose the metrics g_t obtained by shrinking the fibers have constant scalar curvature. If under certain conditions there is an inclusion of (a nontrivial subgroup of) the isometry group of F in the isometry group of (M, g_t), $t \in \,]0, \tau[$, then one can employ the equivariant techniques to deal with possible compensation of eigenvalues and still obtain infinitely many bifurcation values in this nonhomogeneous context.

7. Multiplicity of solutions to the Yamabe problem

We now explain how to obtain the multiplicity result claimed in the Introduction.

Proposition 7.1. *Let g_t, with $t \in]0,\tau[$, be a family of metrics on M with $N(g_t) > 0$ and suppose there exists a sequence $\{t_q\}$ in $]0,\tau[$, that converges to 0, of bifurcation values for g_t. Then, there is an infinite subset $\mathcal{T} \subset]0,\tau[$ accumulating at 0, such that for each $t \in \mathcal{T}$, there are at least 3 solutions to the Yamabe problem in the conformal class $[g_t]$.*

Proof. For all t, denote by $\hat{g}_t$ the unit volume metric homothetic to g_t. Since t_q is a bifurcation value, there are values of t arbitrarily close to t_q for which the conformal class $[g_t]$ contains a unit volume constant scalar curvature metric g distinct from $\hat{g}_t$. Since $N(g_t) > 0$, by continuity of the Morse index, also $N(g) > 0$. In particular, neither $\hat{g}_t$ nor g are minima of the Hilbert–Einstein functional in $[g_t]$. Therefore, $[g_t]$ contains at least 3 distinct unit volume constant scalar curvature metrics, that is, 3 solutions to the Yamabe problem. The set $\mathcal{T}$ of such t's clearly accumulates at 0, since t_q converges to 0. □

Theorems A and B now follow easily from Theorems 4.1 and 6.1, respectively. Indeed, in order to apply Proposition 7.1, it is only necessary to verify that $N(g_t) > 0$. In the first case, if g_t comes from a homogeneous fibration, there must be a value τ of t, when $\mathrm{scal}(t)/(m-1)$ crosses the first eigenvalue from the base, before any compensation is even possible. At this value $t = \tau$, the Morse index changes from 0 to a positive integer. Then, for $t \in]0,\tau[$, we have $N(g_t) \geq N(g_{\tau-\varepsilon}) > 0$. In the second case, $N(g_t)$ gets arbitrarily large as $t \to 0$, so this condition is also satisfied.

References

[Ambrosetti and Malchiodi 1999] A. Ambrosetti and A. Malchiodi, "A multiplicity result for the Yamabe problem on S^n", *J. Funct. Anal.* **168**:2 (1999), 529–561. MR 2000k:53032 Zbl 0949.53028

[Anderson 2005] M. T. Anderson, "On uniqueness and differentiability in the space of Yamabe metrics", *Commun. Contemp. Math.* **7**:3 (2005), 299–310. MR 2006d:53031 Zbl 1082.58013

[Aubin 1976] T. Aubin, "Équations différentielles non linéaires et problème de Yamabe concernant la courbure scalaire", *J. Math. Pures Appl.* (9) **55**:3 (1976), 269–296. MR 55 #4288 Zbl 0336.53033

[Bérard-Bergery 1978] L. Bérard-Bergery, "Sur la courbure des métriques riemanniennes invariantes des groupes de Lie et des espaces homogènes", *Ann. Sci. École Norm. Sup.* (4) **11**:4 (1978), 543–576. MR 80k:53078 Zbl 0426.53038

[Bérard-Bergery and Bourguignon 1982] L. Bérard-Bergery and J.-P. Bourguignon, "Laplacians and Riemannian submersions with totally geodesic fibres", *Illinois J. Math.* **26**:2 (1982), 181–200. MR 84m:58153 Zbl 0483.58021

[Berti and Malchiodi 2001] M. Berti and A. Malchiodi, "Non-compactness and multiplicity results for the Yamabe problem on S^n", *J. Funct. Anal.* **180**:1 (2001), 210–241. MR 2002b:53049 Zbl 0979.53038

[Bessa et al. 2012] G. P. Bessa, J. F. Montenegro, and P. Piccione, "Riemannian submersions with discrete spectrum", *J. Geom. Anal.* **22**:2 (2012), 603–620. MR 2891739 Zbl 1257.58018

[Besse 1987] A. L. Besse, *Einstein manifolds*, Results in Mathematics and Related Areas (3) **10**, Springer, Berlin, 1987. MR 88f:53087 Zbl 0613.53001

[Bettiol and Piccione 2013] R. G. Bettiol and P. Piccione, "Bifurcation and local rigidity of homogeneous solutions to the Yamabe problem on spheres", *Calc. Var. Partial Differential Equations* **47**:3-4 (2013), 789–807. MR 3070564

[Chavel 1984] I. Chavel, *Eigenvalues in Riemannian geometry*, Pure and Applied Mathematics **115**, Academic Press, Orlando, FL, 1984. MR 86g:58140 Zbl 0551.53001

[Fischer and Marsden 1975] A. E. Fischer and J. E. Marsden, "Deformations of the scalar curvature", *Duke Math. J.* **42**:3 (1975), 519–547. MR 52 #1804 Zbl 0336.53032

[Kielhöfer 2004] H. Kielhöfer, *Bifurcation theory: An introduction with applications to PDEs*, Applied Mathematical Sciences **156**, Springer, New York, 2004. MR 2004i:47133 Zbl 1032.35001

[Lima et al. 2012] L. L. de Lima, P. Piccione, and M. Zedda, "On bifurcation of solutions of the Yamabe problem in product manifolds", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **29**:2 (2012), 261–277. MR 2901197 Zbl 1239.58005

[Pollack 1993] D. Pollack, "Nonuniqueness and high energy solutions for a conformally invariant scalar equation", *Comm. Anal. Geom.* **1**:3-4 (1993), 347–414. MR 94m:58051 Zbl 0848.58011

[Schoen 1984] R. Schoen, "Conformal deformation of a Riemannian metric to constant scalar curvature", *J. Differential Geom.* **20**:2 (1984), 479–495. MR 86i:58137 Zbl 0576.53028

[Schoen 1989] R. M. Schoen, "Variational theory for the total scalar curvature functional for Riemannian metrics and related topics", pp. 120–154 in *Topics in calculus of variations* (Montecatini Terme, 1987), edited by M. Giaquinta, Lecture Notes in Math. **1365**, Springer, Berlin, 1989. MR 90g:58023 Zbl 0702.49038

[Schoen 1991] R. M. Schoen, "On the number of constant scalar curvature metrics in a conformal class", pp. 311–320 in *Differential geometry*, edited by B. Lawson and K. Tenenblat, Pitman Monogr. Surveys Pure Appl. Math. **52**, Longman Sci. Tech., Harlow, 1991. MR 94e:53035 Zbl 0733.53021

[Smoller and Wasserman 1990] J. Smoller and A. G. Wasserman, "Bifurcation and symmetry-breaking", *Invent. Math.* **100**:1 (1990), 63–95. MR 91f:58019 Zbl 0721.58011

[Trudinger 1968] N. S. Trudinger, "Remarks concerning the conformal deformation of Riemannian structures on compact manifolds", *Ann. Scuola Norm. Sup. Pisa* (3) **22** (1968), 265–274. MR 39 #2093 Zbl 0159.23801

[Yamabe 1960] H. Yamabe, "On a deformation of Riemannian structures on compact manifolds", *Osaka Math. J.* **12** (1960), 21–37. MR 23 #A2847 Zbl 0096.37201

[Ziller 1982] W. Ziller, "Homogeneous Einstein metrics on spheres and projective spaces", *Math. Ann.* **259**:3 (1982), 351–358. MR 84h:53062 Zbl 0469.53043

Received August 6, 2012.

RENATO G. BETTIOL
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NOTRE DAME
NOTRE DAME, IN 46556-4618
UNITED STATES
rbettiol@nd.edu

PAOLO PICCIONE
DEPARTAMENTO DE MATEMÁTICA
UNIVERSIDADE DE SÃO PAULO
SÃO PAULO, SP, 05508-090
BRAZIL
piccione@ime.usp.br

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.23

RANK GRADIENT OF SMALL COVERS

DARLAN GIRÃO

We prove that if $M \to P$ is a small cover of a compact right-angled hyperbolic polyhedron then M admits a cofinal tower of finite sheeted covers with positive rank gradient. As a corollary, if $\pi_1(M)$ is commensurable with the reflection group of P, then M admits a cofinal tower of finite sheeted covers with positive rank gradient.

1. Introduction

Let P^n be an n-dimensional simple convex polytope. Here P^n is *simple* if the number of codimension-one faces meeting at each vertex is n. Equivalently, the dual K_P of its boundary complex ∂P^n is an $(n-1)$-dimensional simplicial sphere. A *small cover* of P^n is an n-dimensional manifold endowed with an action of the group $\mathbb{Z}_2^n$ whose orbit space is P^n. The notion of small cover was introduced and studied by Davis and Januszkiewicz [1991]. We will be dealing mostly with 3-dimensional polytopes. In the case P is a compact right-angled polyhedron in $\mathbb{H}^3$ then Andreev's theorem [1970] implies that all vertices have valence three and in particular P is a simple convex polytope.

Let G be a finitely generated group. The *rank of G* is the minimal number of elements needed to generate G, and is denoted by $\mathrm{rk}(G)$. If G_j is a finite index subgroup of G, the Reidemeister–Schreier process [Lyndon and Schupp 1977] gives an upper bound on the rank of G_j.

$$\mathrm{rk}(G_j) - 1 \leq [G : G_j](\mathrm{rk}(G) - 1)$$

Lackenby [2006] introduced the notion of *rank gradient*. Given a finitely generated group G and a collection $\{G_j\}$ of finite index subgroups, the rank gradient of the pair $(G, \{G_j\})$ is defined by

$$\mathrm{rgr}(G, \{G_j\}) = \lim_{j\to\infty} \frac{\mathrm{rk}(G_j) - 1}{[G : G_j]}.$$

We say that the collection of finite index subgroups $\{G_j\}$ is *cofinal* if $\cap_j G_j = \{1\}$, and we call it a *tower* if $G_{j+1} < G_j$.

MSC2010: 57M05.
Keywords: rank gradient, hyperbolic 3-manifolds.

In general it is very hard to construct cofinal families $(G, \{G_j\})$ with positive rank gradient. For instance, it seems that only recently the first examples of torsion-free finite covolume Kleinian groups with this property were given in [Girão 2011]. Before stating the main result we need some terminology.

If M is a finite volume hyperbolic 3-manifold, we call the family of covers $\{M_j \to M\}$ *cofinal* (resp. a *tower*) if $\{\pi_1(M_j)\}$ is cofinal (resp. a tower). By the rank gradient of the pair $(M, \{M_j\})$, $\mathrm{rgr}(M, \{M_j\})$, we mean the rank gradient of $(\pi_1(M), \{\pi_1(M_j)\})$.

Theorem 1.1. *Let $M \to P$ be a small cover of a compact, right-angled hyperbolic polyhedron of dimension* 3*. Then M admits a cofinal tower of finite sheeted covers $\{M_j \to M\}$ with positive rank gradient.*

We remark here that this is not true for 3-dimensional polytopes in general. Let $T^3 \to C$ be the covering of a cube in 3-dimensional Euclidean space by the 3-torus T^3. It is easy to see that any subgroup of $\pi_1(T^3) = \mathbb{Z}^3$ has bounded rank and therefore the rank gradient with respect to any tower of covers is zero.

This theorem has the following consequence:

Corollary 1.2. *Let M be a finite volume hyperbolic* 3*-manifold such that $\pi_1(M)$ is commensurable with the group generated by reflections along the faces of a compact, right-angled hyperbolic polyhedron $P \subset \mathbb{H}^3$. Then M admits a cofinal tower of finite sheeted covers $\{M_j \to M\}$ with positive rank gradient.*

We note that this corollary is complementary to the results of [Girão 2011], where ideal right-angled polyhedra were considered. The key idea there was to estimate the rank of the fundamental group of the manifolds by estimating their number of cusps. Here the estimates on the rank of the fundamental groups are given in terms of the rank of the mod 2 homology.

The study of the rank of the fundamental group of (finite volume hyperbolic) 3-manifolds has always been a central theme in low-dimensional topology. In recent years the study of the rank gradient for this class of groups has received special attention. Motivated by the seminal paper [Lackenby 2006], for instance, Long, Lubotzky, Reid proved in [Long et al. 2008] that every finite volume hyperbolic 3-manifold has a cofinal tower of covers in which the *Heegaard genus* grows linearly with the degree of the covers. Whether or not the same happens to the rank of their fundamental groups is a major open problem. Another important recent work using these notions is [Abért and Nikolov 2012]. There they connect the problem related to the growth of the rank of π_1 and the growth of the Heegaard genus in a cofinal tower of hyperbolic 3-manifolds to a problem in topological dynamics, the *fixed price problem* (see [Farber 1998; Gaboriau 2000]). These papers have all been motivation for the current work.

2. Small covers

Recall that an n-dimensional convex polytope P^n is simple if the number of codimension-one faces meeting at each vertex is n. Equivalently, the dual K_P of its boundary complex ∂P is an $(n-1)$-dimensional simplicial complex. A small cover of P is an n-dimensional manifold endowed with an action of the group $\mathbb{Z}_2^n$ whose orbit space is P.

Let K be a finite simplicial complex of dimension $n-1$. For $0 \le i \le n-1$, let f_i be the number of i-simplices of K. Define a polynomial $\Phi_K(t)$ of degree n by

$$\Phi_K(t) = (t-1)^n + \sum_{i=0}^{n-1} f_i (t-1)^{n-1-i}$$

and let h_i be the coefficient of t^{n-i} in this polynomial, that is,

$$\Phi_K(t) = \sum_{i=0}^{n} h_i t^{n-i}.$$

If we restrict to the case where K is the dual K_P of the boundary complex of a convex simple polytope P^n, then one can see that f_i is the number of faces of P^n of codimension $i+1$. Let $h_i(P^n)$ denote the coefficient of t^{n-i} in $\Phi_{K_P}(t)$.

We state one of the main results of [Davis and Januszkiewicz 1991] in our particular setting:

Theorem 2.1. *If $\pi : M^n \to P^n$ is a small cover of a simple convex polytope P^n and $b_i(M^n, \mathbb{Z}_2)$ is the i-th* mod 2 *Betti number of M^n, then $b_i(M^n, \mathbb{Z}_2) = h_i(P^n)$.*

As Davis and Januszkiewicz observe, it is somewhat surprising that all mod 2 Betti numbers of a small cover M^n depend on P^n only. They showed that this theorem does not hold for homology groups in general. They proved that small covers of a square Q by tori and Klein bottles are such that the rational Betti numbers are not determined by Q.

When P is a right-angled dodecahedron in $\mathbb{H}^3$ then [Garrison and Scott 2003] shows that up to homeomorphism there exist exactly 25 small covers of P. Choi [2010] estimates the number of orientable small covers of the n-dimensional cube. Also, if P is a 3-dimensional convex polytope, [Nakayama and Nishimura 2005] proves that P admits an orientable small cover. They also prove that unless P is a 3-simplex, then it admits a nonorientable small cover.

3. Proof of theorem

Theorem 1.1. *Let $M \to P$ be a small cover of a compact, right-angled hyperbolic polyhedron of dimension* 3*. Then M admits a cofinal tower of finite sheeted covers $\{M_j \to M\}$ with positive rank gradient.*

Proof. As observed above, when P is a compact right-angled polyhedron in $\mathbb{H}^3$ then Andreev's theorem [1970] implies that all vertices have valence three and in particular P is a simple convex polytope. Let V, E and F denote the number of vertices, edges and faces, respectively, of a 3-dimensional simple polyhedron P. Straightforward computations show that

$$\Phi_{K_P}(t) = t^3 + (F-3)t^2 + (3-2F+E)t + (V-E+F-1)$$

and thus $h_0(P) = 1$, $h_1(P) = F-3$, $h_2(P) = 3-2F+E$ and $h_3(P) = V-E+F-1$. Since P is simple we also have $E = \frac{3}{2}V$. And since $V - E + F = 2$ (∂P is topologically a sphere) this gives $F = \frac{1}{2}V + 2$ and therefore $h_1(P) = \frac{1}{2}V - 1$.

The strategy involved in the proof is similar to the proof of the main theorem in [Girão 2011]. Given $P \in \mathbb{H}^3$, construct a family of polyhedra

$$P = P_0, P_1, \dots, P_j, \dots$$

such that P_{j+1} is obtained from P_j by reflecting P_j along one of its faces. This must be done in a way such that the following holds: If $x \in \mathbb{H}^3$, then there exists j sufficiently large so that x lies in the interior of P_j. This means that the family $\{P_j\}$ is an exhaustion of $\mathbb{H}^3$. Denote by G_j the group generated by reflections along the faces of P_j. If the family $\{P_j\}$ is constructed as above, then it is easy to see that $G_{j+1} < G_j$ (with index 2) and it can be shown that the tower $\{G_j\}$ is cofinal (see [Agol 2008]). We refer the reader to [Girão 2011] for a detailed proof of this fact.

Now let $M \to P$ be a small cover of P, and let $M_j \to M$ be the cover corresponding to the group $\pi_1(M) \cap G_j$. Recall that the degree of the cover $M \to P$ is 2^3.

Lemma 3.2. $[\pi_1(M_j) : \pi_1(M_{j+1})] = 2$.

Proof of lemma. First observe that $[G_j : G_{j+1}] = 2$. Since $\pi_1(M_1) = G_1 \cap \pi_1(M)$, we must have $[\pi_1(M) : \pi_1(M_1)] \leq 2$. If this index were 1, then it would mean that $\pi_1(M_1) = \pi_1(M) < G_1$ from which would follow that M_1 is a manifold cover of the simple polyhedron P_1 of degree 2^2. But this is not possible, since any manifold cover of a 3-dimensional simple polyhedron must have degree at least 2^3 (see [Davis and Januszkiewicz 1991; Garrison and Scott 2003]). The remaining cases follow by induction. □

Since $[G_j : G_{j+1}] = 2$, from the above lemma and an inductive argument we see that $M_j \to P_j$ is a cover of degree 2^3. In particular this implies that M_j is a small cover of P_j. From Theorem 2.1 we have

$$b_1(M_j, \mathbb{Z}_2) = h_1(P_j).$$

Denote by V_j the number of vertices of P_j. From the computations of h_1,

$$b_1(M_j, \mathbb{Z}_2) = h_1(P_j) = \tfrac{1}{2}V_j - 1.$$

Also note that a lower bound for $\mathrm{rk}(\pi_1(M_j))$ is $b_1(M_j, \mathbb{Z}_2)$ and thus

$$\mathrm{rk}(\pi_1(M_j)) \geq \tfrac{1}{2} V_j - 1.$$

We also have $[\pi_1(M) : \pi_1(M_j)] = 2^j$. Therefore

$$\mathrm{rgr}(\pi_1(M), \{\pi_1(M_j)\}) = \lim_{j\to\infty} \frac{\mathrm{rk}(\pi_1(M_j)) - 1}{[\pi_1(M) : \pi_1(M_j)]} \geq \lim_{j\to\infty} \frac{V_j - 3}{2^{j+1}}.$$

We thus need to show that V_j is of magnitude 2^j. This follows from the next result:

Theorem 3.3 [Atkinson 2009]. *There exist constants $C, D > 0$ such that if P is a compact right-angled polyhedron in $\mathbb{H}^3$ with V vertices then*

$$C(V-8) \leq \mathrm{vol}(P) \leq D(V-10).$$

We now observe that, in our setting, $\mathrm{vol}(P_j) = 2^j \mathrm{vol}(P)$ and thus

$$D(V_j - 10) \geq 2^j \mathrm{vol}(P) \geq 2^j C(V-8)$$

which gives

$$V_j \geq 2^j \frac{C}{D}(V-8) + 10,$$

where V is the number of vertices in P. Also, the second inequality in Atkinson's theorem provides $V > 8$. □

4. Extending the examples

Theorem 1.1 has an interesting corollary, which complements the family of manifolds provided in [Girão 2011].

Corollary 1.2. *Let N be a closed hyperbolic 3-manifold such that $\pi_1(N)$ is commensurable with the group generated by reflections along the faces of a compact, right-angled hyperbolic polyhedron $P \subset \mathbb{H}^3$. Then N admits a cofinal tower of finite sheeted covers $\{N_j \to N\}$ with positive rank gradient.*

Proof. First we note that, by passing to a finite cover, we may assume N is orientable. Note also that [Nakayama and Nishimura 2005] implies orientable small covers of P exist and therefore N is commensurable with a small cover $M \to P$. Let N' be the manifold cover of both M and N corresponding to the group $\pi_1(M) \cap \pi_1(N)$. Consider now $N_j \to N$ corresponding to the group $\pi_1(N') \cap G_j$, where the family $\{G_j\}$ is given as in the proof of Theorem 1.1. Consider also $\{M_j\}$, the tower where M_j is a small cover of P_j, as in the proof of Theorem 1.1.

Note that $\pi_1(N_j) = \pi_1(N') \cap G_j = \pi_1(N') \cap \pi_1(M_j) < \pi_1(M_j)$ and therefore we have the following diagram of covers, where the labels in the arrows indicate the degree of the cover:

$$\begin{tikzcd}
 & P \arrow[l, phantom] & P_1 \arrow[l, "2"'] & \cdots \arrow[l, "2"'] & P_j \arrow[l, "2"'] & \cdots \arrow[l, "2"'] \\
 & M \arrow[u, "2^3"'] & M_1 \arrow[l, "2"'] \arrow[u, "2^3"'] & \cdots \arrow[l, "2"'] & M_j \arrow[l, "2"'] \arrow[u, "2^3"'] & \cdots \arrow[l, "2"'] \\
N & N' \arrow[l] \arrow[u] & N_1 \arrow[l] \arrow[u] & \cdots \arrow[l] & N_j \arrow[l] \arrow[u] & \cdots \arrow[l]
\end{tikzcd}$$

Agol, Culler and Shalen proved:

Theorem 4.2 ([Agol et al. 2006]; see also [Shalen 2007]). *Let M be a closed, orientable hyperbolic 3-manifold such that $b_1(M, \mathbb{Z}_p) = r$ for a given prime p. Then for any finite sheeted covering space M' of M, $b_1(M', \mathbb{Z}_p) \geq r - 1$.*

We thus have

$$\operatorname{rk}(\pi_1(N_j)) \geq b_1(N_j, \mathbb{Z}_2) \geq b_1(M_j, \mathbb{Z}_2) - 1 = \tfrac{1}{2}V_j - 2$$

and therefore all we need to do is show that $[\pi_1(N) : \pi_1(N_j)]$ grows at most as fast as 2^j. But from the above diagram we see that $[\pi_1(N_j) : \pi_1(N_{j+1})] \leq 2$ and we are done. $\square$

Acknowledgements

The author thanks Alan Reid for suggesting he look at [Davis and Januszkiewicz 1991]. He also thanks the Erwin Schrödinger International Institute for Mathematical Physics for the hospitality and financial support during his visit in the occasion of the workshop "Golod–Shafarevich Groups and Algebras and Rank Gradient". Finally, he would like to thank the organizers of this event.

References

[Abért and Nikolov 2012] M. Abért and N. Nikolov, "Rank gradient, cost of groups and the rank versus Heegaard genus problem", *J. Eur. Math. Soc.* **14**:5 (2012), 1657–1677. MR 2966663 Zbl 06095894

[Agol 2008] I. Agol, "Criteria for virtual fibering", *J. Topol.* **1**:2 (2008), 269–284. MR 2009b:57033 Zbl 1148.57023

[Agol et al. 2006] I. Agol, M. Culler, and P. B. Shalen, "Dehn surgery, homology and hyperbolic volume", *Algebr. Geom. Topol.* **6** (2006), 2297–2312. MR 2008f:57024 Zbl 1129.57019

[Andreev 1970] E. M. Andreev, "Convex polyhedra in Lobachevski spaces", *Mat. Sb.* (*N.S.*) **81 (123)** (1970), 445–478. In Russian; translated in *Math. USSR Sbornik* **10**:3 (1970), 413–440. MR 41 #4367 Zbl 0217.46801

[Atkinson 2009] C. K. Atkinson, "Volume estimates for equiangular hyperbolic Coxeter polyhedra", *Algebr. Geom. Topol.* **9**:2 (2009), 1225–1254. MR 2010k:57035 Zbl 1170.57012

[Choi 2010] S. Choi, "The number of orientable small covers over cubes", *Proc. Japan Acad. Ser. A Math. Sci.* **86**:6 (2010), 97–100. MR 2011i:57046 Zbl 1198.37074

[Davis and Januszkiewicz 1991] M. W. Davis and T. Januszkiewicz, "Convex polytopes, Coxeter orbifolds and torus actions", *Duke Math. J.* **62**:2 (1991), 417–451. MR 92i:52012 Zbl 0733.52006

[Farber 1998] M. Farber, "Geometry of growth: Approximation theorems for L^2 invariants", *Math. Ann.* **311**:2 (1998), 335–375. MR 2000b:58042 Zbl 0911.53026

[Gaboriau 2000] D. Gaboriau, "Coût des relations d'équivalence et des groupes", *Invent. Math.* **139**:1 (2000), 41–98. MR 2001f:28030 Zbl 0939.28012

[Garrison and Scott 2003] A. Garrison and R. Scott, "Small covers of the dodecahedron and the 120-cell", *Proc. Amer. Math. Soc.* **131**:3 (2003), 963–971. MR 2003h:57018 Zbl 1009.57019

[Girão 2011] D. Girão, "Rank gradient in cofinal towers of certain Kleinian groups", preprint, 2011. To appear in *Groups, Geometry, and Dynamics*. arXiv 1102.4281

[Lackenby 2006] M. Lackenby, "Heegaard splittings, the virtually Haken conjecture and property (τ)", *Invent. Math.* **164**:2 (2006), 317–359. MR 2007c:57030 Zbl 1110.57015

[Long et al. 2008] D. D. Long, A. Lubotzky, and A. W. Reid, "Heegaard genus and property τ for hyperbolic 3-manifolds", *J. Topol.* **1**:1 (2008), 152–158. MR 2008j:57036 Zbl 1158.57018

[Lyndon and Schupp 1977] R. C. Lyndon and P. E. Schupp, *Combinatorial group theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete **89**, Springer, Berlin, 1977. MR 58 #28182 Zbl 0368.20023

[Nakayama and Nishimura 2005] H. Nakayama and Y. Nishimura, "The orientability of small covers and coloring simple polytopes", *Osaka J. Math.* **42**:1 (2005), 243–256. MR 2006a:57023 Zbl 1065.05041

[Shalen 2007] P. B. Shalen, "Hyperbolic volume, Heegaard genus and ranks of groups", pp. 335–349 in *Workshop on Heegaard splittings* (Haifa, 2005), edited by C. Gordon and Y. Moriah, Geom. Topol. Monogr. **12**, Geometry & Topology Publications, Coventry, 2007. MR 2009k:57029 Zbl 1140.57009

Received September 3, 2012. Revised January 2, 2013.

DARLAN GIRÃO
DEPARTMENT OF MATHEMATICS
UNIVERSIDADE FEDERAL DO CEARÁ
AV. HUMBERTO MONTE S/N
CAMPUS DO PICI - BLOCO 914
60455-760 FORTALEZA, CE
BRAZIL

dgirao@mat.ufc.br

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.31

NONRATIONALITY OF NODAL QUARTIC THREEFOLDS

KYUSIK HONG

We prove the factoriality of every nodal quartic threefold with 13 singular points that contains neither planes nor quadric surfaces. As a corollary, any nodal quartic threefold with 13 singular points that contains neither planes nor quadric surfaces is nonrational.

1. Introduction

All varieties are assumed to be projective, normal, and defined over $\mathbb{C}$. A nodal variety is one that has, at most, isolated ordinary double points (nodes). A variety V is said to be factorial if each Weil divisor is Cartier, and $\mathbb{Q}$-factorial if a multiple of each Weil divisor of V is Cartier. This simple-looking definition is quite subtle when applied to projective varieties. It depends both on the kind of singularities and on their position. In the case of a Fano threefold X, $\mathbb{Q}$-factoriality is equivalent to the condition $\operatorname{rank}(H^2(X, \mathbb{Z})) = \operatorname{rank}(H_4(X, \mathbb{Z}))$. Thus a smooth Fano threefold is always $\mathbb{Q}$-factorial. The local class group at a node in a threefold has no torsion [Milnor 1968], so each Weil divisor that is $\mathbb{Q}$-Cartier must be a Cartier divisor on a nodal hypersurface in $\mathbb{P}^4$. Therefore, $\mathbb{Q}$-factoriality is equivalent to factoriality for a nodal hypersurface in $\mathbb{P}^4$. Moreover, the factoriality of a nodal quartic threefold implies its nonrationality; Mella [2004] proved that every factorial nodal quartic threefold is nonrational. This generalizes a classical result by Iskovskikh and Manin [1971] that every smooth quartic threefold is nonrational. On the other hand, there exist nonfactorial nodal quartic threefolds that are nonrational.

In view of Mella's result and the importance of rationality, studying the factoriality of nodal quartic threefolds is of interest. Here we consider this problem when the number of nodes is 13. This extends earlier results, which we now quote.

Throughout, X_4 will represent a nodal quartic threefold.

Theorem 1.1 [Cheltsov 2006]. *A quartic X_4 with at most 9 nodes is factorial if it contains no plane.*

MSC2010: 14C20, 14J17, 14J30.
Keywords: factoriality, rationality, nodal quartic threefold.

Theorem 1.2 [Shramov 2007, Theorem 1.3]. *A quartic X_4 with at most 11 nodes is factorial if it contains no planes. If X_4 has 12 nodes, then X_4 is factorial, with the exception of the case when X_4 contains a quadric surface.*

In Section 3 of this paper, we prove the following results.

Theorem 1.3. *A quartic X_4 with at most 13 nodes is factorial if it contains neither planes nor quadric surfaces.*

Corollary 1.4. *A quartic X_4 with at most 13 nodes is nonrational if it contains neither planes nor quadric surfaces.*

Theorem 1.3 improves on the degree-4 case of [Cheltsov 2006, Conjecture 13], which generalizes a well-known conjecture by Ciliberto [2004].

We present an example which motivates our study.

Example 1.5 [Cheltsov 2006, Example 10]. Let a_2, h_2, b_3 and g_1 be homogeneous polynomials of degrees 2, 2, 3 and 1, respectively. Consider the quartic threefold X_4 defined by the equation

$$a_2(x, y, z, w, t)h_2(x, y, z, w, t) = b_3(x, y, z, w, t)g_1(x, y, z, w, t);$$

it is the general quartic threefold passing through the quadric surface Q defined by $a_2 = g_1 = 0$. The quartic X_4 has 12 nodes, given by $h_2 = g_1 = a_2 = b_3 = 0$, and it is not factorial.

2. Preliminaries

If the nodes of a nodal quartic threefold X_4 impose independent linear conditions on hypersurfaces of degree 3 in $\mathbb{P}^4$, then X_4 is factorial [Cynk 2001]. Hence, if a nodal quartic threefold X_4 is factorial, it must have at most 35 simple double points because $h^0(\mathbb{P}^4, \mathbb{O}_{\mathbb{P}^4}(3)) = 35$. A nodal quartic threefold X_4 cannot have more than 45 nodes [Friedman 1986; Varchenko 1983]. Moreover, there is a unique nodal quartic threefold with 45 nodes [de Jong et al. 1990]. It is known as the Burkhardt quartic, which has too many nodes to be factorial.

The following result is one of the main tools.

Theorem 2.1 [Eisenbud and Koh 1989, Theorem 2]. *Let Σ be a set of points in $\mathbb{P}^N$ and let $d \geq 2$ be an integer. If no $dk+2$ of the points of Σ lie in a projective k-plane for all $k \geq 1$, then Σ imposes independent conditions on forms of degree d in $\mathbb{P}^N$.*

We see that the singular points of nodal threefolds are located in $\mathbb{P}^4$ with the following properties:

Lemma 2.2 [Cheltsov and Park 2006, Lemma 2.9]. *Let X_d be a nodal hypersurface of degree d in $\mathbb{P}^4$.*

(1) *A curve of degree k in $\mathbb{P}^4$ contains at most $k(d-1)$ nodes of X_d.*

(2) *If a 2-plane contains $d(d-1)/2+1$ nodes of X_d, then the 2-plane is contained in X_d.*

More generally, a nodal hypersurface X_d of degree d in $\mathbb{P}^4$ is factorial if and only if the singular points of X_d impose linearly independent conditions on hypersurfaces of degree $2d-5$ in $\mathbb{P}^4$ [Cynk 2001]. To prove the factoriality of X_d, we have to compute whether $h^1(\mathbb{P}^4, \mathbb{O}_{\mathbb{P}^4}(2d-5)\otimes \mathcal{I}_{\mathrm{Sing}(X_d)})$ is 0 or not. Therefore, we need to study the dimension of the linear system $|\mathbb{O}_{\mathbb{P}^4}(2d-5)|$ having assigned base points.

Fixing a nodal hypersurface $X_d\subset\mathbb{P}^4$ and r singular points of X_d in $\mathbb{P}^4$, what is the dimension of the space of hypersurfaces of degree $2d-5$ in $\mathbb{P}^4$ passing through those points?

Let $\phi:\tilde{\mathbb{P}}^4\to\mathbb{P}^4$ be the blowing up of $\mathbb{P}^4$ along $\mathrm{Sing}(X_d)=\{p_1,\dots,p_r\}$. Let $\tilde{X}_d$ be the strict transform of X_d, let H be the divisor class of the pullback of a hyperplane under ϕ, and let $E=\sum_{i=1}^r E_i$, where the E_i are the classes of exceptional divisors. Suppose that $\mathcal{F}\in\mathrm{Pic}\,\mathbb{P}^4$ and $\mathcal{F}'=\phi^*\mathcal{F}$. Then $\phi_*\mathbb{O}_{\tilde{\mathbb{P}}^4}\cong\mathbb{O}_{\mathbb{P}^4}$ and $R^i\phi_*\mathcal{F}'=0$ for $i>0$. Therefore, $H^j(\mathbb{P}^4,\mathcal{F})\cong H^j(\tilde{\mathbb{P}}^4,\mathcal{F}')$ by the Leray spectral sequence. Moreover, we have the equalities

$$\phi_*(\mathbb{O}_{\tilde{\mathbb{P}}^4}(-E))=\mathcal{I}_{\mathrm{Sing}(X_d)},\quad R^i\phi_*(\mathbb{O}_{\tilde{\mathbb{P}}^4}(-E))=0\quad\text{for } i>0,\tag{2.3}$$

$$\phi_*(\mathbb{O}_{\tilde{\mathbb{P}}^4}(kE))=\mathbb{O}_{\mathbb{P}^4},\quad R^i\phi_*(\mathbb{O}_{\tilde{\mathbb{P}}^4}(kE))=0\ \text{ for } i>0,\ k=0,1,2,3.\tag{2.4}$$

By (2.3), we get $h^j(\tilde{\mathbb{P}}^4,\mathbb{O}_{\tilde{\mathbb{P}}^4}((2d-5)H-E))=h^j(\mathbb{P}^4,\mathbb{O}_{\mathbb{P}^4}(2d-5)\otimes\mathcal{I}_{\mathrm{Sing}(X_d)})$.

Let $L_4(2d-5;1^r)=L_4(2d-5;1_1,1_2,\dots,1_r)$ be the complete linear system $\left|(2d-5)H-\sum_{i=1}^r E_i\right|$ on $\tilde{\mathbb{P}}^4$. We will use the same notation to denote the corresponding line bundle on $\tilde{\mathbb{P}}^4$, as well as the push-forward of $\left|(2d-5)H-\sum_{i=1}^r E_i\right|$ to $\mathbb{P}^4$, i.e., the linear system of threefolds of degree $2d-5$ with multiplicity 1 at p_i.

Definition 2.5. A nonempty linear system $L_4(2d-5;1^r)$ is special if

$$h^0(\tilde{\mathbb{P}}^4,L_4(2d-5;1^r))>(2d-1)(2d-2)(2d-3)(2d-4)/24-r$$

or, which is the same, if $h^1(\tilde{\mathbb{P}}^4,L_4(2d-5;1^r))\neq 0$.

Note that $h^0(\tilde{\mathbb{P}}^4,L_4(2d-5;1^r))\geq h^0(\mathbb{P}^4,\mathbb{O}_{\mathbb{P}^4}(2d-5))-r$. We call the system $L_4(2d-5;1^r)$ nonspecial if $h^0(\tilde{\mathbb{P}}^4,L_4(2d-5;1^r))=h^0(\mathbb{P}^4,\mathbb{O}_{\mathbb{P}^4}(2d-5))-r$, or, which is the same, if $h^1(\tilde{\mathbb{P}}^4,L_4(2d-5;1^r))=0$.

Lemma 2.6. $h^1(\tilde{X}_d,\mathbb{O}_{\tilde{X}_d}((2d-5)H-E))=0\Leftrightarrow h^1(\tilde{\mathbb{P}}^4,\mathbb{O}_{\tilde{\mathbb{P}}^4}((2d-5)H-E))=0$

Proof. Consider the exact sequence

$$0\to\mathbb{O}_{\tilde{\mathbb{P}}^4}(-\tilde{X}_d+(2d-5)H-E)\to\mathbb{O}_{\tilde{\mathbb{P}}^4}((2d-5)H-E)\to\mathbb{O}_{\tilde{X}_d}((2d-5)H-E)\to 0.$$

We have $\tilde{X}_d \equiv dH - 2E$. By (2.4),

$$R^i\phi_*(\mathcal{O}_{\tilde{\mathbb{P}}^4}((d-5)H+E)) = \mathcal{O}_{\mathbb{P}^4}(d-5)\otimes R^i\phi_*(\mathcal{O}_{\tilde{\mathbb{P}}^4}(E)) = 0$$

for $i > 0$. Then we get $h^j(\tilde{\mathbb{P}}^4, \mathcal{O}_{\tilde{\mathbb{P}}^3}((d-5)H+E)) = h^j(\mathbb{P}^4, \mathcal{O}_{\mathbb{P}^4}(d-5)) = 0$ for $0 < j < 4$. Thus

$$h^1(\tilde{X}_d, \mathcal{O}_{\tilde{X}_d}((2d-5)H-E)) = h^1(\tilde{\mathbb{P}}^4, \mathcal{O}_{\tilde{\mathbb{P}}^4}((2d-5)H-E)). \qquad \square$$

Therefore, studying linear systems of threefolds with assigned base points p_i is equivalent to studying complete linear systems on the fourfold $\tilde{\mathbb{P}}^4$ obtained by blowing up the points p_i. Also, a nodal hypersurface X_d of degree d in $\mathbb{P}^4$ is factorial if and only if a nonempty linear system $L_4(2d-5;\, 1^r)$ is nonspecial.

In the rest of this section we present tools to investigate the speciality of $L_4(3;\, 1^r)$ for $r \geq 12$. We need to consider the restriction on a quadric surface Q due to Example 1.5 and Lemma 2.2.

Before stating these results, let Q be a smooth quadric surface (when Q is a singular quadric, we don't have a proof yet). Let $\mathrm{Sing}(X_d)\cap Q = \{p_1, p_2, \ldots, p_\lambda\}$, where λ the maximal number of points of $\mathrm{Sing}(X_d)$ that can belong to the quadric Q. We consider a linear system $|\mathcal{O}(k_1,k_2)\otimes\mathcal{I}_{\mathrm{Sing}(X_d)\cap Q}| := L_Q((k_1,k_2);\, 1^\lambda)$ with $k_1 > 0$ and $k_2 > 0$ on the quadric Q (that is, a system of curves of type (k_1,k_2) through points p_i of multiplicity 1). Using a similar method to Definition 2.5, we define the speciality for $L_Q(k_1,k_2)$. Then we see that

$$h^0(Q, L_Q((k_1,k_2);\, 1^\lambda)) = k_1k_2 + k_1 + k_2 + 1 - \lambda$$

if and only if the system $L_Q((k_1,k_2);\, 1^\lambda)$ is nonspecial.

To prove the factoriality of a nodal quartic X_4, we have to investigate the speciality of the restriction system $L_4(3;\, 1^r)|_Q = L_Q((3,3);\, 1^\lambda)$.

Lemma 2.7. *With the above notation, let* $\#|\mathrm{Sing}(X_4)| = r \geq 12$. *Suppose the smooth quadric surface Q is defined by $\{f_2(x,y,z,w,t) = 0\}\cap A_1$, where f_2 is a homogeneous polynomial of degree* 2 *and A_1 is a hyperplane in $\mathbb{P}^4$ such that* $\#|A_1\cap\mathrm{Sing}(X_4)| \geq 12$. *Let λ be the maximal number of points of* $\mathrm{Sing}(X_4)$ *that can belong to the smooth quadric Q.*

(1) *Suppose that $0 \leq r-\lambda \leq 2$. Then a linear system $L_Q((3,3);\, 1^\lambda)$ is special if and only if a linear system* $L_4(3;\, 1^r)$ *is special.*

(2) *Suppose that $3 \leq r-\lambda$. If a linear system $L_Q((3,3);\, 1^\lambda)$ is special, then a linear system* $L_4(3;\, 1^r)$ *is special.*

Proof. Let $A_1 \equiv H - \widehat{E}$ and let $\{f_2 = 0\} \equiv 2H - \widehat{E}$, where $\widehat{E} = \sum_{i=1}^{\lambda} E_i$.

Consider the exact sequence

$$0 \to \mathcal{O}_{\tilde{\mathbb{P}}^4}(-H+\widehat{E}+3H-E) \to \mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E) \to \mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E)|_{A_1} \to 0.$$

We get the exact sequence

$$
\begin{aligned}
(2.8)\quad 0 \to H^0\big(\mathcal{O}_{\tilde{\mathbb{P}}^4}\big(2H-\sum_{i=\lambda+1}^{r} E_i\big)\big) &\to H^0(\mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E)) \\
&\to H^0(\mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E)|_{A_1}) \to H^1\big(\mathcal{O}_{\tilde{\mathbb{P}}^4}\big(2H-\sum_{i=\lambda+1}^{r} E_i\big)\big) \\
&\to H^1(\mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E)) \to H^1(\mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E)|_{A_1}) \to 0.
\end{aligned}
$$

Notice that $R^j\phi_*\mathcal{O}_{\tilde{\mathbb{P}}^4}(2H-E)=\mathcal{O}_{\mathbb{P}^4}(2)\otimes R^j\phi_*\mathcal{O}_{\tilde{\mathbb{P}}^4}(-E)=0$ for all $j>0$. If $r=\lambda$, then $h^1\big(\mathcal{O}_{\tilde{\mathbb{P}}^4}\big(2H-\sum_{i=\lambda+1}^{r} E_i\big)\big)=h^1(\mathcal{O}_{\mathbb{P}^4}(2))=0$. Theorem 2.1 and Lemma 2.2(1) tell us that $h^1\big(\mathcal{O}_{\tilde{\mathbb{P}}^4}\big(2H-\sum_{i=\lambda+1}^{r} E_i\big)\big)=h^1(\mathbb{P}^4,\mathcal{I}_{\sum_{i=\lambda+1}^{r} p_i}(2))=0$ for $1\le r-\lambda\le 5$. We then have $h^1(L_4(3;1^r))=h^1(\mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E))=h^1(\mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E)|_{A_1})$ when $0\le r-\lambda\le 5$.

Also, note that $Q\equiv(2H-\widehat{E})|_{A_1}$. From the short exact sequence

$$0\to\mathcal{O}_{A_1}(-2H+\widehat{E}+3H-E)\to\mathcal{O}_{A_1}(3H-E)\to\mathcal{O}_{A_1}(3H-E)|_Q\to 0,$$

we obtain the sequence

$$
\begin{aligned}
(2.9)\quad 0 \to H^0\big(\mathcal{O}_{A_1}\big(H-\sum_{i=\lambda+1}^{r} E_i\big)\big) &\to H^0(\mathcal{O}_{A_1}(3H-E)) \\
&\to H^0(\mathcal{O}_{A_1}(3H-E)|_Q) \to H^1\big(\mathcal{O}_{A_1}\big(H-\sum_{i=\lambda+1}^{r} E_i\big)\big) \\
&\to H^1(\mathcal{O}_{A_1}(3H-E)) \to H^1(\mathcal{O}_{A_1}(3H-E)|_Q) \to 0.
\end{aligned}
$$

Note that $h^1\big(\mathcal{O}_{A_1}\big(H-\sum_{i=\lambda+1}^{r} E_i\big)\big)=h^1(\mathbb{P}^3,\mathcal{I}_{\sum_{i=\lambda+1}^{r} p_i}(1))=0$ for $0\le r-\lambda\le 2$. We have $h^1(\mathcal{O}_{A_1}(3H-E))=h^1(\mathcal{O}_{A_1}(3H-E)|_Q)=h^1(L_Q((3,3);1^\lambda))$ when $0\le r-\lambda\le 2$.

The second statement follows from the last lines of (2.8) and (2.9). □

Corollary 2.10. *With the above notation, if $r=\lambda+j$ for $j=0,1,2$, then*

$$
\begin{aligned}
&h^0(\mathcal{O}_{\tilde{\mathbb{P}}^4}(3H-E)) \\
&\quad = h^0\big(L_Q((3,3);1^\lambda)\big)+h^0\big(\mathcal{O}_{A_1}\big(H-\sum_{i=\lambda+1}^{r} E_i\big)\big)+h^0\big(\mathcal{O}_{\tilde{\mathbb{P}}^4}\big(2H-\sum_{i=\lambda+1}^{r} E_i\big)\big).
\end{aligned}
$$

Proof. This follows immediately from Lemma 2.7(1). □

Lemma 2.11. *Suppose that a curve of type (a,b), with $0<a\le b\le 3$, and a curve of type $(3,3)$ meet in $3a+3b$ points, say $\Sigma_{(a,b)}=\{p_1,\dots,p_{3a+3b}\}$.*

(1) *Let $\Psi\subset\Sigma_{(2,2)}$ and let $\alpha=\#|\Psi|$. Then a linear system $L_Q((3,3);1^\alpha)$ is special if and only if $\alpha=12$, i.e., a curve of type $(2,2)$ contains $\Sigma_{(2,2)}$.*

(2) *Let* $\Omega \subset \Sigma_{(2,3)}$ (*or* $\Sigma_{(3,2)}$) *and let* $\beta = \#|\Omega|$. *Then a linear system* $L_Q((3,3); 1^\beta)$ *is special if and only if* $\beta = 14$.

(3) *Let* $\Upsilon \subset \Sigma_{(3,3)}$ *and let* $\gamma = \#|\Upsilon|$. *Then a linear system* $L_Q((3,3); 1^\gamma)$ *is special if and only if* $\gamma = 15$.

Proof. By Lemma 2.2(1), a system $L_Q((3,3); 1^\lambda)$ has no fixed curve. The number $h_{\Sigma_{(a,b)}}(3,3)$ of conditions imposed by $\Sigma_{(a,b)}$ on forms of bidegree $(3,3)$ satisfies

$$h_{\Sigma_{(a,b)}}(3,3) = h^0(L_Q(3,3)) - h^0(L_Q(3-a, 3-b)) - 1.$$

There are four possible cases for the speciality of $L_Q((3,3); 1^\lambda)$.

When $(a,b) = (2,2)$, we get $h_{\Sigma_{(2,2)}}(3,3) = 11 < \#|\Sigma_{(2,2)}|$.

When $(a,b) = (2,3)$ (or $(3,2)$), we can write $h_{\Sigma_{(2,3)}}(3,3) = 13 < \#|\Sigma_{(2,3)}|$ (or $\#|\Sigma_{(3,2)}|$), so the statement (2) is true.

Finally, the inequality $h_{\Sigma_{(3,3)}}(3,3) = 14 < \#|\Sigma_{(3,3)}|$ implies statement (3). □

3. The proof of Theorem 1.3

Let X_4 be a nodal hypersurface in $\mathbb{P}^4$.

Definition 3.1. The set $\mathrm{Sing}(X_4)$ satisfies the property ∇ if the following conditions hold:

- There is a hyperplane A_1 in $\mathbb{P}^4$ which contains at least 11 points of $\mathrm{Sing}(X_4)$.
- Fix an arbitrary point p of $A_1 \cap \mathrm{Sing}(X_4)$. There is a reducible cubic surface in A_1 passing through $(A_1 \cap \mathrm{Sing}(X_4)) \setminus \{p\}$ but not passing through p.

Lemma 3.2. *Let* $\#|\mathrm{Sing}(X_4)| = 11$. *Suppose that there is a hyperplane* A_1 *in* $\mathbb{P}^4$ *such that* $A_1 \cap \mathrm{Sing}(X_4) = \mathrm{Sing}(X_4)$, *and every quadric surface in* A_1 *does not contain all the points of* $\mathrm{Sing}(X_4)$. *Then* $\mathrm{Sing}(X_4)$ *satisfies the property* ∇.

Proof. Fix an arbitrary point p of $\mathrm{Sing}(X_4)$. Let $\mathrm{Sing}(X_4) = \{p_1, p_2, \dots, p_{10}, p\}$. Since every quadric surface does not contain all the points of $\mathrm{Sing}(X_4)$, we can find a quadric surface Q_1 in A_1 containing 9 points, say $\{p_1, p_2, \dots, p_8, p_9\}$, of $\mathrm{Sing}(X_4) \setminus \{p\}$ but not containing p. We shall take for the required cubic surface the union of Q_1 and a two-dimensional linear subspace in A_1 passing through p_{10} and not passing through p. □

Lemma 3.3. *Let* $\mathcal{M} \subseteq |\mathcal{O}_{\mathbb{P}^3}(2)|$ *be a linear subsystem that contains the set* $A_1 \cap \mathrm{Sing}(X_4)$, *where* A_1 *is a hyperplane in* $\mathbb{P}^4$. *Suppose that* $n = \#|A_1 \cap \mathrm{Sing}(X_4)| \geq 11$, X_4 *contains no* 2*-planes*, *and a space curve of degree* 4 *in* A_1 *contains at most* 10 *points of* $\mathrm{Sing}(X_4)$. *Then the base locus* $Bs(\mathcal{M})$ *is empty or two-dimensional.*

Proof. Suppose that $Bs(\mathcal{M})$ is zero-dimensional. Let M_1, M_2, and M_3 be the general surfaces of $\mathcal{M}$. Then the intersection number $M_1 \cdot M_2 \cdot M_3$ has at most 8, but $n \geqslant 11$ holds.

Now we suppose that the curve $B = Bs(\mathcal{M}) \subset A_1$. Then $\deg B \leq 4$. Since $n \geqslant 11$, by Lemma 2.2(1), $\deg B = 4$, and B must be reduced. Moreover, B is not contained in a two-dimensional linear subspace, because a two-dimensional linear subspace contains at most 6 points. This contradicts the assumption. □

Lemma 3.4. *Let* $\#|\mathrm{Sing}(X_4)| = 11$. *Suppose that* X_4 *contains no* 2*-planes, there is a hyperplane* A_1 *in* $\mathbb{P}^4$ *such that* $A_1 \cap \mathrm{Sing}(X_4) = \mathrm{Sing}(X_4)$, *every reducible quadric surface does not contain all the points of* $\mathrm{Sing}(X_4)$, *and a space curve of degree* 4 *in* A_1 *does not pass through all the points of* $\mathrm{Sing}(X_4)$. *Then* $\mathrm{Sing}(X_4)$ *satisfies the property* ∇.

Proof. Fix an arbitrary point p of $\mathrm{Sing}(X_4)$. Let $\mathrm{Sing}(X_4) = \{p_1, p_2, \ldots, p_{10}, p\}$. By Lemma 3.2, we assume that there is an irreducible quadric surface Q_2 in A_1 containing all the points of $\mathrm{Sing}(X_4)$. By Lemma 3.3, any quadric surface in A_1 passing through all the points of $\mathrm{Sing}(X_4)$ coincides with Q_2. Suppose that Q_2 is determined by 8 points, say $\{p_1, p_2, \ldots, p_8\}$, of $\mathrm{Sing}(X_4) \setminus \{p\}$ together with p. Then we can find a quadric Q_3 in A_1 containing $\{p_1, p_2, \ldots, p_8\}$ and not containing p. We can assume that $p_k \notin Q_3$ for $k = 9$ or 10; otherwise, take a two-dimensional linear subspace in A_1 containing the point $\mathrm{Sing}(X_4) \setminus Q_3$ but not containing p.

If $p \notin \overline{p_9, p_{10}}$, then we can easily construct a reducible cubic surface in A_1 that contains $\mathrm{Sing}(X_4) \setminus \{p\}$ and does not contain p.

Now we suppose that three points $\{p_9, p_{10}, p\}$ lie on a single line. By statement (1) of Lemma 2.2, the line determined by $\{p_i, p_9\}$ (or $\{p_i, p_{10}\}$), for $1 \leq i \leq 8$, does not pass through p. We consider the quadric surface Q_4 determined by $\{p_1, p_2, \ldots, p_8, p_9\}$. We can assume that the quadric surface Q_4 contains the points p; otherwise, take a two-dimensional linear subspace in A_1 containing the point p_{10} but not containing p.

Then Q_4 must be Q_2, that is, the quadric surface Q_2 is determined by the point p and $\{p_1, p_2, \ldots, p_8, p_9\} \setminus \{p_j\}$ for $1 \leq j \leq 9$. Therefore, we can find a quadric surface Q_5 passing through $\{p_1, p_2, \ldots, p_8, p_9\} \setminus \{p_j\}$, for $1 \leq j \leq 8$ and not passing through p. Let l be the line determined by two points $\mathrm{Sing}(X_4) \setminus Q_5 \setminus \{p\}$. Then p cannot lie on the line l. Let $\bar{A}_1$ be a two-dimensional linear subspace in A_1 containing the line l but not containing p. Then the union of Q_5 and $\bar{A}_1$ is the desired form of degree 3. □

Lemma 3.5. *Let* $\#|\mathrm{Sing}(X_4)| = 11$. *Suppose that* X_4 *contains no* 2*-planes, there is a hyperplane* A_1 *in* $\mathbb{P}^4$ *such that* $A_1 \cap \mathrm{Sing}(X_4) = \mathrm{Sing}(X_4)$, *every reducible quadric surface does not contain all the points of* $\mathrm{Sing}(X_4)$, *and there is a space* *curve D of degree* 4 *in A_1 that passes through all the points of* $\mathrm{Sing}(X_4)$. *Then* $\mathrm{Sing}(X_4)$ *satisfies the property* ∇.

Proof. Fix an arbitrary point p of $\mathrm{Sing}(X_4)$. Let $\mathrm{Sing}(X_4) = \{p_1, p_2, \dots, p_{10}, p\}$. By Lemma 3.2, we assume that there is an irreducible quadric surface containing all the points of $\mathrm{Sing}(X_4)$. By Lemma 2.2(1), a twisted cubic contains at most 9 singular points of X_4. Lemma 3.3 tells us that D can be written as an intersection of two different quadric surfaces in A_1. Then there is a quadric surface in A_1 containing 7 points, say $\{p_1, \dots, p_7\}$, of $\mathrm{Sing}(X_4) \setminus \{p\}$ but not containing p. We consider the two-dimensional linear subspace $\hat{A}_1$ in A_1 determined by $\{p_8, p_9, p_{10}\}$. We can assume that $p \in \hat{A}_1$; otherwise, one can easily construct the required cubic surface in A_1.

By Lemma 2.2(1), renumbering p_8, p_9, and p_{10} if necessary, we can assume that $\{p_9, p_{10}, p\}$ span $\hat{A}_1$. By Lemma 2.2(2), $\hat{A}_1$ contains at most 2 points of $\{p_1, \dots, p_7\}$. We can assume that $\hat{A}_1$ contains 2 points, say $\{p_6, p_7\}$, of $\{p_1, \dots, p_7\}$ (a similar method applies to the case when $\hat{A}_1$ passes through one or none of $\{p_1, \dots, p_7\}$). Assume that all the quadric surfaces in A_1 containing 7 points, $\{p_1, \dots, p_7, p_8\} \setminus \{p_i\}$ for $1 \le i \le 5$, also pass through p. Then each quadric surface containing the points p_6, p_7, and p_8 also contains p, and hence p_6, p_7, p_8, and p lie on a single line. This is a contradiction to Lemma 2.2(1). Thus, we can find a quadric surface Q_6 in A_1 containing 7 points, $\{p_1, \dots, p_7, p_8\} \setminus \{p_i\}$ for $1 \le i \le 5$, and not containing p. Take the two-dimensional linear subspace A_1' in A_1 determined by three points $\mathrm{Sing}(X_4) \setminus Q_6 \setminus \{p\}$. Then A_1' does not contain p, and hence $Q_6 + A_1'$ is the required cubic surface. □

Proposition 3.6. *Let* $\#|\mathrm{Sing}(X_4)| = 11$. *Suppose that X_4 contains no* 2*-planes, a hyperplane in* $\mathbb{P}^4$ *contains all the points of* $\mathrm{Sing}(X_4)$, *and every reducible quadric surface does not contain all the points of* $\mathrm{Sing}(X_4)$. *Then* $\mathrm{Sing}(X_4)$ *satisfies the property* ∇.

Proof. By Lemma 3.2, we can assume that there is an irreducible quadric surface containing all the points of $\mathrm{Sing}(X_4)$. It immediately follows from Lemmas 3.4 and 3.5. □

The following result is proved in [Shramov 2007].

Corollary 3.7 [Shramov 2007, Corollary 3.4]. *Let* $\mathrm{Sing}(X_4) = \{p_1, \dots, p_s\}$. *Assume that* $s \ge 11$, *and no* 2*-plane contains* 7 *points of* $\mathrm{Sing}(X_4)$. *Let* $p_1, \dots, p_s$ *be points lying in a reducible quadric surface* (*that is, in a pair of planes spanning a three-dimensional surface*). *Then either* $p_1, \dots, p_s$ *impose independent conditions on the forms of degree* 3 *or* $p_1, \dots, p_s$ *also lie in an irreducible quadric surface.*

Proposition 3.8. *Let* $\#|\mathrm{Sing}(X_4)| = 13$. *Suppose that X_4 contains no* 2*-planes, and a hyperplane in* $\mathbb{P}^4$ *contains at most* 11 *points of* $\mathrm{Sing}(X_4)$. *Then X_4 is factorial.*

Proof. Fix an arbitrary point p of $\mathrm{Sing}(X_4)$. It is enough to construct a cubic threefold T that contains $\mathrm{Sing}(X_4) \setminus \{p\}$ and does not contain p.

Suppose that there is a hyperplane A_1 in $\mathbb{P}^4$ containing 11 points of $\mathrm{Sing}(X_4)$; otherwise, X_4 is factorial due to Theorem 2.1 and Lemma 2.2.

Suppose that $p \notin A_1$. We can find a quadric threefold G passing through two points $\mathrm{Sing}(X_4) \setminus A_1 \setminus \{p\}$ but not passing through p. Then $T = A_1 + G$.

Now assume that $p \in A_1$. We divide the case into two subcases. Let $\{q_1, q_2\} = \mathrm{Sing}(X_4) \setminus A_1$.

First, assume that a reducible quadric surface in A_1 does not contain all the points of $\mathrm{Sing}(X_4)$. By Proposition 3.6, we obtain a cubic surface $W = Q' + \Lambda$, where Q' is a quadric surface and Λ is a two-dimensional linear subspace, in A_1 containing $\mathrm{Sing}(X_4) \cap (A_1 \setminus \{p\})$ but not containing p. Take the cone Q'' over Q' with vertex q_1 and a hyperplane Λ' in $\mathbb{P}^4$ containing Λ together with q_2. Then we can get the required cubic threefold T as the union of Q'' and Λ'.

Second, assume that a reducible quadric surface in A_1 contains all the points of $\mathrm{Sing}(X_4)$. Applying the proof of Corollary 3.7, we can construct a reducible cubic surface Y in A_1 passing through $(\mathrm{Sing}(X_4) \cap A_1) \setminus \{p\}$ and not passing through p such that Y consists of three two-dimensional linear subspaces, say L_1, L_2, and L_3. Note that $q_1, q_2 \notin A_1$. Then we obtain the required cubic threefold T as the union of a hyperplane in $\mathbb{P}^4$ containing $\{L_1, q_1\}$, a hyperplane in $\mathbb{P}^4$ containing $\{L_2, q_2\}$, and a hyperplane in $\mathbb{P}^4$ containing L_3 but not containing p. □

The following three results are proved in [Shramov 2007]. They are very useful for the proof of Theorem 1.3. We let $\mathrm{Sing}(X_4) = \{p_1, \dots, p_s\}$.

Lemma 3.9 [Shramov 2007, Lemma 3.5]. *Assume that $s \leq 12$, and no 2-plane contains 7 points of $\mathrm{Sing}(X_4)$. Let $p_1, \dots, p_s$ be points in a 3-dimensional subspace $\mathbb{P}^3 \subset \mathbb{P}^4$ not lying in some quadric surface. Then $p_1, \dots, p_s$ impose independent conditions on the forms of degree 3 in $\mathbb{P}^3$ (and therefore also in $\mathbb{P}^4$).*

Lemma 3.10 [Shramov 2007, Lemma 3.8]. *Assume that no 2-plane contains 7 points of $\mathrm{Sing}(X_4)$. Let $p_1, \dots, p_{12}$ be points in a quadric surface. Then either $p_1, \dots, p_{12}$ impose independent conditions on the forms of degree 3 in $\mathbb{P}^4$ or $p_1, \dots, p_{12}$ lie in a pencil of quadric surfaces in some 3-dimensional subspace.*

Lemma 3.11 [Shramov 2007, Lemma 3.9]. *Assume that no 2-plane contains 7 points of $\mathrm{Sing}(X_4)$. Let $p_1, \dots, p_{12}$ be points lying in a pencil of quadric surfaces in a 3-dimensional subspace. Then X_4 contains a quadric surface.*

Proposition 3.12. *Let $\#|\mathrm{Sing}(X_4)| = 13$. Suppose that X_4 contains no 2-planes, and a hyperplane in $\mathbb{P}^4$ contains at most 12 points of $\mathrm{Sing}(X_4)$. Then X_4 is either factorial or contains a quadric surface.*

Proof. Fix an arbitrary point p of $\mathrm{Sing}(X_4)$. We can assume that there is a hyperplane A_1 in $\mathbb{P}^4$ containing 12 points of $\mathrm{Sing}(X_4)$. Let $\{q\} = \mathrm{Sing}(X_4) \setminus A_1$.

First, suppose that a quadric surface contains at most 11 points of $\mathrm{Sing}(X_4)$. We can assume that $p \in A_1$; otherwise, one can easily check that X_4 is factorial. By Lemma 3.9, we can find a cubic surface U in A_1 passing through $(\mathrm{Sing}(X_4) \cap A_1) \setminus \{p\}$ and not passing through p. Taking a cone over U with vertex q, we obtain a cubic threefold passing through $\mathrm{Sing}(X_4) \setminus \{p\}$ and not passing through p. In this case, X_4 is factorial.

Second, suppose that there is a quadric surface $\widehat{Q}$ containing 12 points, say Ξ, of $\mathrm{Sing}(X_4)$. We can assume that Ξ cannot lie on a pencil of quadric surface in A_1; otherwise, by Lemma 3.11, X_4 contains a quadric surface.

Now we have to prove that X_4 is factorial. We can assume that $p \in \widehat{Q}$. Let $\Xi = \{p_1, \dots, p_{11}, p\}$. Applying the proof of Lemma 3.10, we obtain a reducible cubic surface K in A_1 containing $\Xi \setminus \{p\}$ but not containing p. Note that $q \notin A_1$. Let $K = S + L$, where S is a quadric surface and L is a two-dimensional linear subspace. Then we can construct a cubic threefold as the union of the cone over S with vertex q and a hyperplane in $\mathbb{P}^4$ containing L but not containing p. Thus, X_4 is factorial. $\square$

Proposition 3.13. *Let* $\#|\mathrm{Sing}(X_4)| = 13$. *Suppose that* X_4 *contains no* 2*-planes, and there is a hyperplane* A_1 *in* $\mathbb{P}^4$ *containing all the points of* $\mathrm{Sing}(X_4)$. *Then* X_4 *is either factorial or contains a quadric surface.*

Proof. Fix an arbitrary point p of $\mathrm{Sing}(X_4)$. Let $\mathrm{Sing}(X_4) = \{p_1, \dots, p_{12}, p\}$.

Suppose that every quadric surface does not contain all the points of $\mathrm{Sing}(X_4)$. Then we find a quadric surface containing 9 points, say $\{p_1, \dots, p_9\}$, of $\mathrm{Sing}(X_4)$ but not containing p. We consider the two-dimensional linear subspace $\hat{A}_1$ in A_1 determined by $\{p_{10}, p_{11}, p_{12}\}$. We can assume that $p \in \hat{A}_1$; otherwise, one can easily check that X_4 is factorial.

By Lemma 2.2(1), renumbering p_{10}, p_{11}, and p_{12} if necessary, we can assume that $\{p_{11}, p_{12}, p\}$ span $\hat{A}_1$. By Lemma 2.2(2), $\hat{A}_1$ contains at most 2 points of $\{p_1, \dots, p_9\}$. We can assume that $\hat{A}_1$ contains 2 points, say $\{p_8, p_9\}$, of $\{p_1, \dots, p_9\}$ (a similar method applies to the case when $\hat{A}_1$ passes through one or none of $\{p_1, \dots, p_9\}$). Assume that all the quadric surfaces in A_1 containing 9 points, $\{p_1, \dots, p_9, p_{10}\} \setminus \{p_i\}$ for $1 \leq i \leq 7$, also pass through p. Then each quadric surface containing the points p_8, p_9, and p_{10} also contains p, and hence p_8, p_9, p_{10}, and p lie on a single line. This is a contradiction to Lemma 2.2(1). Thus, we can find a quadric surface Q_7 in A_1 containing 9 points, $\{p_1, \dots, p_9, p_{10}\} \setminus \{p_i\}$ for $1 \leq i \leq 7$, and not containing p. The union of Q_7 and the two-dimensional linear subspace A_1'' in A_1 determined by three points $\mathrm{Sing}(X_4) \setminus Q_7 \setminus \{p\}$ is a cubic surface containing $\mathrm{Sing}(X_4) \setminus \{p\}$ but not containing p. It implies that X_4 is factorial.

Now suppose that there is a quadric surface Q_8 containing all the points of $\mathrm{Sing}(X_4)$. Then, by Lemma 2.2(2), Q_8 is irreducible. We may assume that Q_8 is a quadric cone; otherwise, by Lemma 2.7(1) and 2.11(1), X_4 is either factorial or contains a quadric surface. For instance, since a curve of type (1, 1) contains at most 6 points of $\mathrm{Sing}(X_4)$, X_4 cannot contain a 2-plane.

If there is a quadric surface different from Q_8 containing 12 points of $\mathrm{Sing}(X_4)$, then X_4 contains a quadric surface due to Lemma 3.11.

We can assume Q_8 is unique; that is, any quadric surface different from Q_8 passes through at most 11 points of $\mathrm{Sing}(X_4)$. Consider a nodal quartic threefold $\widehat{X}_4$ defined by

$$a_1(x, y, z, t, w)h_3(x, y, z, t, w) + b_2(x, y, z, t, w)g_2(x, y, z, t, w) = 0,$$

where a_1, h_3, b_2 and g_2 are homogeneous polynomials of degree 1, 3, 2, and 2, respectively. Suppose that the quadric Q_8 is the quadric cone given by $\{a_1 = b_2 = 0\}$, and V_2 is a quadric surface given by $\{a_1 = g_2 = 0\}$. Then the nodes of $\widehat{X}_4$ are $\{a_1 = h_3 = b_2 = g_2 = 0\}$ and the vertex of Q_8. The quartic $\widehat{X}_4$ has 13 nodes with $\mathrm{Sing}(\widehat{X}_4) = \mathrm{Sing}(X_4)$, and all the points of $\mathrm{Sing}(\widehat{X}_4)$ lie on a hyperplane $\{a_1 = 0\}$. By the uniqueness of Q_8, X_4 must be $\widehat{X}_4$. Since V_2 contains 12 points of $\mathrm{Sing}(X_4)$, this contradicts the assumption. □

Proof of Theorem 1.3. Suppose that every two dimensional linear subspace contains at most 6 singular points of a nodal quartic X_4, i.e, by Lemma 2.2(2), X_4 contains no 2-planes; otherwise, X_4 is defined by an equation of the form

$$y_1(x, y, z, t, w) f_3(x, y, z, t, w) + \hat{y}_1(x, y, z, t, w)g_3(x, y, z, t, w) = 0,$$

where y_1, f_3, $\hat{y}_1$, and g_3 are homogeneous polynomials of degree 1, 3, 1, and 3, respectively. Then X_4 is not factorial.

Theorem 1.3 immediately follows from Propositions 3.8, 3.12, and 3.13. □

Acknowledgments

The author would like to thank two referees for their valuable comments and suggestions. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2011-0001565).

References

[Cheltsov 2006] I. Cheltsov, "Nonrational nodal quartic threefolds", *Pacific J. Math.* **226**:1 (2006), 65–81. MR 2007e:14024 Zbl 1123.14010

[Cheltsov and Park 2006] I. Cheltsov and J. Park, "Factorial hypersurfaces in $\mathbb{P}^4$ with nodes", *Geom. Dedicata* **121** (2006), 205–219. MR 2007m:14064 Zbl 1110.14031

[Ciliberto and Di Gennaro 2004] C. Ciliberto and V. Di Gennaro, "Factoriality of certain hypersurfaces of $\mathbb{P}^4$ with ordinary double points", pp. 1–7 in *Algebraic transformation groups and algebraic varieties*, edited by V. L. Popov, Encyclopaedia Math. Sci. **132**, Springer, Berlin, 2004. MR 2005i:14043 Zbl 1068.14010

[Cynk 2001] S. Cynk, "Defect of a nodal hypersurface", *Manuscripta Math.* **104**:3 (2001), 325–331. MR 2002g:14056 Zbl 0983.14017

[Eisenbud and Koh 1989] D. Eisenbud and J.-H. Koh, "Remarks on points in a projective space", pp. 157–172 in *Commutative algebra* (Berkeley, CA, 1987), edited by M. Hochster et al., Math. Sci. Res. Inst. Publ. **15**, Springer, New York, 1989. MR 90i:14008 Zbl 0736.14022

[Friedman 1986] R. Friedman, "Simultaneous resolution of threefold double points", *Math. Ann.* **274**:4 (1986), 671–689. MR 87k:32035 Zbl 0576.14013

[Iskovskih and Manin 1971] V. A. Iskovskih and J. I. Manin, "Three-dimensional quartics and counterexamples to the Lüroth problem", *Mat. Sb.* (*N.S.*) **86(128)**:1(9) (1971), 140–166. In Russian; translated in *Math. USSR-Sb.* **15**:1 (1971), 141–166. MR 45 #266 Zbl 0222.14009

[de Jong et al. 1990] A. J. de Jong, N. I. Shepherd-Barron, and A. Van de Ven, "On the Burkhardt quartic", *Math. Ann.* **286**:1-3 (1990), 309–328. MR 91f:14038 Zbl 0712.14023

[Mella 2004] M. Mella, "Birational geometry of quartic 3-folds, II: The importance of being $\mathbb{Q}$-factorial", *Math. Ann.* **330**:1 (2004), 107–126. MR 2005h:14030 Zbl 1058.14022

[Milnor 1968] J. Milnor, *Singular points of complex hypersurfaces*, Annals of Mathematics Studies **61**, Princeton University Press, 1968. MR 39 #969 Zbl 0184.48405

[Shramov 2007] K. A. Shramov, "$\mathbb{Q}$-factorial quartic threefolds", *Mat. Sb.* **198**:7 (2007), 103–114. In Russian; translated in *Sb. Math.* **198**:8 (2007), 1165–1174. MR 2008i:14062 Zbl 1132.14039

[Varchenko 1983] A. N. Varchenko, "Semicontinuity of the spectrum and an upper bound for the number of singular points of the projective hypersurface", *Dokl. Akad. Nauk SSSR* **270**:6 (1983), 1294–1297. In Russian, translated in *Sov. Math. Dokl.* **27** (1983), 735–739. MR 85d:32028 Zbl 0537.14003

Received April 29, 2012. Revised November 14, 2012.

KYUSIK HONG
DEPARTMENT OF MATHEMATICS AND INSTITUTE FOR MATHEMATICAL SCIENCE
KONKUK UNIVERSITY
1 HWAYANG-DONG
GWANGJIN-GU, SEOUL 143-701
REPUBLIC OF KOREA

kszoo@postech.ac.kr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.43

SUPERTROPICAL LINEAR ALGEBRA

ZUR IZHAKIAN, MANFRED KNEBUSCH AND LOUIS ROWEN

The objective of this paper is to lay out the algebraic theory of supertropical vector spaces and linear algebra, utilizing the key antisymmetric relation of "ghost surpasses". Special attention is paid to the various notions of "base", which include d-base and s-base, and these are compared to other treatments in the tropical theory. Whereas the number of elements in various d-bases may differ, it is shown that when an s-base exists, it is unique up to permutation and multiplication by scalars, and can be identified with a set of "critical" elements. Then we turn to orthogonality of vectors, which leads to supertropical bilinear forms and a supertropical version of the Gram matrix, including its connection to linear dependence. We also obtain a supertropical version of a theorem of Artin, which says that if g-orthogonality is a symmetric relation, then the underlying bilinear form is (supertropically) symmetric.

1. Introduction

The objective of this paper is to lay the foundations for an algebraic theory of linear algebra in supertropical mathematics. Special attention is paid to the notion of base, which plays a subtler role here than in classical linear algebra. Although an extensive literature already exists on tropical linear algebra over the max-plus algebra, including matrix rank [Akian et al. 2006] and linear dependence [Akian et al. 2009], the emphasis often is combinatoric or geometric. The traditional approach in semiring theory is to divide the determinant into a positive and negative part (since -1 need not exist in the semiring); see [Akian et al. 1990]. The ensuing reliance

The research of Izhakian and Rowen is supported by the Israel Science Foundation (grant no. 448/09). Research on this paper was carried out by the three authors through the Research in Pairs program of the Mathematisches Forschungsinstitut Oberwolfach (MFO), Germany, July 2009. The research of Izhakian was supported by the Oberwolfach Leibniz Fellows Programme, MFO. Knebusch was supported in part by the Gelbart Institute at Bar-Ilan University, the Minerva Foundation at Tel-Aviv University, the Mathematics Department of Bar-Ilan University, and the Emmy Noether Institute. The authors are grateful to Tal Perri for helpful comments on an earlier version.

MSC2010: primary 11D09, 15A03, 15A04, 15A15, 15A63, 16Y60; secondary 51M20, 20M18, 15A33, 14T05.

Keywords: tropical algebra, supertropical vector spaces, linear algebra, change of base semirings, linear and bilinear forms, Gram matrix.

on combinatorics leads to competing (and different) definitions. For example, in [Akian et al. 2006], five different definitions of matrix rank are given: the row rank, the Barvinok (Shein) rank, the strong rank, the Gondran–Minoux rank, the symmetrized rank, and the Kapranov rank. Since [Maclagan and Sturmfels 2009] is an excellent source of fundamental results and examples, we use it as a general reference for the "standard" tropical theory and compare several of our notions with the definitions given there.

Extending the max-plus algebra to the algebra of [Izhakian and Rowen 2010] yields a theory paralleling the classical structure theory of commutative algebras. The structure theory of supertropical predomains tends to unify these notions, giving a single formula for the determinant, from which we can define a nonsingular matrix; then the row rank, column rank, and strong rank all coincide. This makes it easier to proceed with a traditional algebraic development. Explicit properties of matrices, especially singularity, were studied in [Izhakian and Rowen 2009; 2011a; 2011b; 2011c]. We elaborate a bit, by proving in Theorem 3.5 that the product of two nonsingular $n \times n$ matrices cannot be in $M_n(\mathcal{G}_{\mathbb{0}})$ (although it might be singular).

Our main objective in this paper is to study bases of supertropical vector spaces over supertropical semifields, relying as far as possible on the structure theory. While this theory parallels the classical theory of linear algebra, several key differences do emerge. The major difference is that whereas in classical linear algebra, any maximal independent set is a minimal spanning set, these two notions differ significantly in the tropical theory. Thus, we must consider two kinds of "base", one defined by means of independence and the other defined by means of spanning.

First, one can take a maximal (tropically) independent set, which we call a *d-base*, called a "basis" in [Maclagan and Sturmfels 2009, Definition 5.2.4]. This has considerable geometric significance, intuitively providing a notion of rank (although, by an example in [Maclagan and Sturmfels 2009], the rank might vary according to the choice of d-base). As one might expect from [Izhakian and Rowen 2011b], any dependence among vectors can be extended to an (often unique) *saturated dependence*, which is maximal in a certain sense; see Theorem 4.20. This leads to a delicate analysis of the rank of a subspace, especially since it turns out that the number of elements in different d-bases may differ.

Alternatively, one may consider sets that (tropically) span the subspace; a minimal such set, when it exists, is called an *s-base*. Such sets are used in generating convex spaces, as studied in [Feichtner and Sturmfels 2005]. Not every d-base is an s-base, and not every s-base is a d-base. In fact, the number of elements of an s-base might necessarily be larger than the number of elements of a d-base. Surprisingly, an s-base is unique up to scalar multiples, by Corollary 5.25, and actually can be characterized in terms of *critical* elements, which intuitively are elements that cannot be decomposed into sums of other elements; see Theorem 5.24. On the other hand, the d-bases can be quite varied, and lead us to interesting subspaces that they span.

In the last section, we introduce supertropical bilinear forms in order to study “left ghost orthogonality” between vectors. One calls two vectors v and w *left g-orthogonal* with respect to a supertropical bilinear form $\langle\ ,\ \rangle$ when $\langle v, w\rangle$ is a ghost. We construct the *Gram matrix* and prove the connection between tropical dependence of vectors in a nondegenerate space and the singularity of their Gram matrix (Theorem 6.7). Finally, we prove Theorem 6.19, a variant of Artin's theorem: When the g-orthogonality relation is symmetric in the sense of Definition 6.12, the supertropical bilinear form is “supertropically symmetric”.

2. Supertropical structures

Supertropical semifields. A *semiring without zero*, which we notate as a *semiring*†, is a structure $(R, +, \cdot, \mathbb{1}_R)$ such that $(R, \cdot, \mathbb{1}_R)$ is a monoid and $(R, +)$ is a commutative semigroup, with distributivity of multiplication over addition on both sides. (In other words, a semiring† does not necessarily have the zero element $\mathbb{0}$, but any semiring can also be considered as a semiring†.) Given a semiring† $R^\dagger$, we can formally adjoin the element $\mathbb{0}$ to obtain the semiring $R := R^\dagger \cup \{\mathbb{0}\}$, where we stipulate, for all $a \in R$,

$$\mathbb{0} + a = a + \mathbb{0} = a, \qquad \mathbb{0}a = a\mathbb{0} = \mathbb{0}.$$

A *supertropical semiring*† is a triple $(R^\dagger, \mathcal{G}, \nu)$, where $R^\dagger$ is a semiring† and $\mathcal{G}$ is a semiring† ideal called the *ghost ideal*, together with an idempotent map

$$\nu : R^\dagger \to \mathcal{G}$$

(preserving multiplication as well as addition) called the *ghost map on* $R^\dagger$, satisfying the properties

(a) $a + b = a^\nu$ if $a^\nu = b^\nu$, and

(b) $a + b \in \{a, b\}$ for all $a, b \in R^\dagger$ such that $a^\nu \neq b^\nu$. (Equivalently, $\mathcal{G}$ is ordered via $a^\nu \leq b^\nu$ if and only if $a^\nu + b^\nu = b^\nu$.)

In particular, $\nu(a) = a + a$. We write a^ν for $\nu(a)$.

Remark 2.1. The element $e := \nu(\mathbb{1}_R)$ is both a multiplicative and additive idempotent of $R^\dagger$, which plays a key role, since $\nu(R^\dagger) = eR^\dagger$. Each element of $eR^\dagger$ is additively idempotent:

$$er + er = (e + e)r = er.$$

We write $a >_\nu b$ if $a^\nu > b^\nu$.

Definition 2.2. Elements a and b of $(R^\dagger, \mathcal{G}, \nu)$ are *ν-matched*, written $a \cong_\nu b$, if $a^\nu = b^\nu$. The element a *dominates* b if $a >_\nu b$.

Recall that any commutative supertropical semiring satisfies the *Frobenius formula* from [Izhakian and Rowen 2010, Remark 1.1]:

$$(a+b)^m = a^m + b^m \tag{2-1}$$

for any $m \in \mathbb{N}^+$.

A *supertropical semifield*† is a supertropical semiring† $(F^\dagger, \mathcal{G}, \nu)$ in which $\mathcal{T} := F^\dagger \setminus \mathcal{G}$ is an Abelian group and the restriction from ν to $\mathcal{T}$ maps onto $\mathcal{G}$. $\mathcal{T}$ is called the set of *tangible elements* of $F^\dagger$.

We have the analogous definitions when we adjoin the element $\mathbb{0}_F$ to the supertropical semifield† $F^\dagger$ to obtain the *supertropical semifield* F. Thus, we write

$$F := F^\dagger \cup \{\mathbb{0}_F\} = (F, \mathcal{G}_\mathbb{0}, \nu),$$

where $\mathcal{G}_\mathbb{0} := \mathcal{G} \cup \{\mathbb{0}_F\}$ is the *ghost ideal* and the ghost map $\nu : F \to \mathcal{G}_\mathbb{0}$ satisfies $\nu(\mathbb{0}_F) = \mathbb{0}_F$. Conversely, given a supertropical semifield $(F, \mathcal{G}_\mathbb{0}, \nu)$, we can take $F^\dagger = F \setminus \{\mathbb{0}_F\}$ and $\mathcal{G} = \mathcal{G}_\mathbb{0} \setminus \{\mathbb{0}_F\}$ and define the supertropical semifield† $(F^\dagger, \mathcal{G}, \nu)$. Thus, the theories with or without $\mathbb{0}_F$ are basically the same. *Throughout the remainder of this paper, $F = (F, \mathcal{G}_\mathbb{0}, \nu)$ denotes a supertropical semifield.*

Intuitively, the tangible elements correspond in some sense to the original max-plus algebra, although here $a + a = a^\nu$ instead of $a + a = a$. Our motivating example [Izhakian 2009] of a supertropical semifield, used as the primary example throughout [Izhakian and Rowen 2010] as well as in this paper, is the extended tropical semiring

$$\mathbb{T} := D(\mathbb{R}) := \big(\mathbb{R} \cup \mathbb{R}^\nu \cup \{-\infty\}, \mathbb{R}^\nu \cup \{-\infty\}, 1_\mathbb{R}\big),$$

whose operations are induced by the standard operations max and $+$ over the real numbers; we call this *logarithmic notation*, since the zero element $\mathbb{0}_\mathbb{T}$ is $-\infty$ and the unit element $\mathbb{1}_\mathbb{T}$ is 0. To clarify our exposition, most of the examples in this paper are presented for $D(\mathbb{R})$. On the other hand, it is often convenient to take $\mathbb{Q}$ or $\mathbb{Z}$ instead of $\mathbb{R}$, especially when working with powers of elements, and the "characteristic 1" semifield† consisting of the single element $\mathbb{1}$ has interesting applications. (This becomes the well-known "Boolean semifield" $\{\mathbb{0}, \mathbb{1}\}$ when the zero element is adjoined.) Semirings of polynomials over supertropical semifields also play a crucial role in tropical geometry. Nonidempotent semifields† such as $\mathbb{Q}^+$ have applications to arithmetic. Accordingly, our theory is framed for vector spaces over supertropical semifields†, although, in analogy to the classical theory, the major theorems could also be formulated over semidomains.

The supertropical semifield plays a basic role in supertropical algebra parallel to the role of the field in classical algebra. Accordingly, one is led to study linear algebra over supertropical semifields.

Occasionally, we also want to pass back from $\mathcal{G}$ to $\mathcal{T}$. Abusing notation slightly, we pick a representative in $\mathcal{T}$ for each class in the image of $\widehat{\nu}$, thereby getting a function $\widehat{\nu} : R^{\dagger} \to \mathcal{T}$ by putting $\widehat{\nu}|_{\mathcal{T}} = 1_{\mathcal{T}}$.

The "ghost surpass" and "ghost dependence" relations. We consider the supertropical semiring $(R, \mathcal{G}_{\mathbb{0}}, \nu)$.

Definition 2.3. We say b is *ghost dependent* on a, written $b \curlyvee_{\mathrm{gd}} a$, if $a+b \in \mathcal{G}_{\mathbb{0}}$.

In particular, $a \cong_{\nu} b$ implies that $a \curlyvee_{\mathrm{gd}} b$.

The ghost dependence relation is symmetric, but not transitive, since $1 \curlyvee_{\mathrm{gd}} 3^{\nu}$ and $3^{\nu} \curlyvee_{\mathrm{gd}} 2$, although 1 and 2 are not ghost dependent. The following antisymmetric and transitive relation is a key to much of the theory.

Definition 2.4. We define the *ghost surpasses* relation $\models_{\mathrm{gs}}$ on a supertropical semiring $R = (R, \mathcal{G}_{\mathbb{0}}, \nu)$ by

$$a \models_{\mathrm{gs}} b \iff a = b + c \text{ for some } c \in \mathcal{G}_{\mathbb{0}}.$$

In this notation, by writing $a \models_{\mathrm{gs}} \mathbb{0}_R$ we mean $a \in \mathcal{G}_{\mathbb{0}}$. This restricts to the *ghost surpasses* relation on $R^{\dagger}$ by

$$a \models_{\mathrm{gs}} b \iff a = b \text{ or } a = b + c \text{ for some } c \in \mathcal{G}.$$

Remark 2.5. The following are equivalent:

(1) $a \curlyvee_{\mathrm{gd}} \mathbb{0}_R$.

(2) $a \in \mathcal{G}_{\mathbb{0}}$.

(3) $a \models_{\mathrm{gs}} \mathbb{0}_R$.

We quote some easy properties of $\models_{\mathrm{gs}}$ from [Izhakian and Rowen 2011b]:

Remark 2.6. (i) When a is tangible, $a \models_{\mathrm{gs}} b$ implies $a = b$ [loc. cit., Remark 1.2]. In particular, tangible elements are comparable under $\models_{\mathrm{gs}}$ if and only if they are equal. In this way, the relation $\models_{\mathrm{gs}}$ generalizes equality.

(ii) $a \models_{\mathrm{gs}} b$ if and only if $a = b$ or a is a ghost $\geq_{\nu} b$. In particular, if $a \models_{\mathrm{gs}} b$ then $a \geq_{\nu} b$; if $a \models_{\mathrm{gs}} b$ for $b \in \mathcal{G}_{\mathbb{0}}$, then $a \in \mathcal{G}_{\mathbb{0}}$.

(iii) The relation $\models_{\mathrm{gs}}$ is a partial order on R [loc. cit., Lemma 1.5].

(iv) If $a \models_{\mathrm{gs}} b$, then $a \curlyvee_{\mathrm{gd}} b$.

Supertropical vector spaces. Modules over semirings (often called "semimodules" in the literature, or sometimes "cones") are defined just as modules over rings, except that now the additive structure is that of a semigroup instead of a group. (Subtraction does not enter into the other axioms of a module over a ring.)

Definition 2.7. Let R be a semiring. An *R-module* V is a semigroup $(V, +, \mathbb{0}_V)$ together with scalar multiplication $R \times V \to V$ satisfying the following properties for all $r_i \in R$ and $v, w \in V$:

(1) $r(v+w) = rv + rw$.

(2) $(r_1 + r_2)v = r_1 v + r_2 v$.

(3) $(r_1 r_2)v = r_1(r_2 v)$.

(4) $\mathbb{1}_R v = v$.

(5) $r\mathbb{0}_V = \mathbb{0}_V$.

(6) $\mathbb{0}_R v = \mathbb{0}_V$.

Note 2.8. One could also define a module over a semiring† by deleting Axiom (6). In the other direction, any module V over a semiring† $R^\dagger$ becomes an R-module when we formally define $\mathbb{0}_R v = \mathbb{0}_V$ for each $v \in V$.

The reason we prefer the terminology "module" is that this definition of module over a semiring R coincides with the usual definition of module when R is a ring, since $-v = (-\mathbb{1}_R)v$.

In case the underlying semiring R is a supertropical semifield F, V is called a *(supertropical) vector space* over F, or vector space for short. We focus on vector spaces in this paper, and call their elements *vectors*. Our main example of a vector space in this paper, as well as the main example in [Izhakian and Rowen 2011a], is $F^{(n)}$, whose ghost map acts as ν on each component. The zero element $\mathbb{0}$ of $F^{(n)}$ is $(\mathbb{0}_F, \dots, \mathbb{0}_F)$. However, we also are interested in subspaces of $F^{(n)}$, so we need this greater generality.

A supertropical vector space V has the distinguished *standard ghost submodule* $\mathcal{H}_{\mathbb{0}} := eV$, as well as the *ghost map* $\nu : V \to \mathcal{H}_{\mathbb{0}}$, given by $\nu(v) := v + v = ev$. We write v^ν for $\nu(v)$. For example, $F^{(n)}$ has the standard ghost submodule $\mathcal{G}_{\mathbb{0}}^{(n)}$, with

$$(v_1, \dots, v_n)^\nu = (v_1^\nu, \dots, v_n^\nu).$$

Definition 2.9. The *tangible vectors* of $F^{(n)}$ are those $(v_1, \dots, v_n) \neq \mathbb{0}$ such that each $v_i \in \mathcal{T}_{\mathbb{0}}$.

Lemma 2.10. *The following properties are satisfied for all $\alpha \in F$, $v, w \in F^{(n)}$:*

(1) $(\alpha v)^\nu = \alpha v^\nu = \alpha^\nu v$.

(2) $(v+w)^\nu = v^\nu + w^\nu$.

Proof. (1) $(\alpha v)^\nu = e(\alpha v) = (e\alpha)v = (\alpha e)v = \alpha(ev) = \alpha v^\nu$.

(2) $(v+w)^\nu = e(v+w) = ev + ew = v^\nu + w^\nu$. $\square$

As with supertropical semirings, we define the *ghost surpassing relation* $\models_{\mathrm{gs}}$ for vectors $v = (v_1, \dots, v_n)$, $w = (w_1, \dots, w_n)$ by

$$v \models_{\mathrm{gs}} w \quad \text{if} \quad v_i \models_{\mathrm{gs}} w_i \text{ for } 1 \le i \le n.$$

We say that two vectors v, w are ν-*matched*, written $v \cong_\nu w$, if $v^\nu = w^\nu$. Likewise, we write $v \ge_\nu w$ if $v^\nu \models_{\mathrm{gs}} w^\nu$.

Example 2.11.

$$(v_1, \dots, v_n) \ge_\nu (w_1, \dots, w_n) \tag{2-2}$$

in $R^{(n)}$ if and only if $v_i \ge_\nu w_i$ for each $1 \le i \le n$.

Also, for elements v, w in V, we define

$$v \curlyvee_{\mathrm{gd}} w \quad \text{if} \quad v + w \in \mathcal{H}_{\mathbb{0}}.$$

Remark 2.12. (i) If $v \models_{\mathrm{gs}} w$, then $v + w \in \mathcal{H}_{\mathbb{0}}$; thus $v \curlyvee_{\mathrm{gd}} w$.

(ii) If $v_i \models_{\mathrm{gs}} w$ for $i = 1, 2$, then $v_1 + v_2 \models_{\mathrm{gs}} w$.

Lemma 2.13. *The following property holds for all* $v, w \in V$, $h \in \mathcal{H}_{\mathbb{0}}$:

$$v = w + h \implies v + h = v.$$

Proof. $v = w + h = w + h + h = v + h$. ☐

Proposition 2.14. *Any vector space* V *satisfies the following property, for* v, h_1, $h_2 \in V$:

$$v + h_1 + h_2 = v \implies v + h_2 = v.$$

Proof. $v = v + h_1 + h_2 = (v + h_1 + h_2) + (h_1 + h_2) = v + eh_1 + eh_2$. Take $w = v + eh_1$ and $h = eh_2$ in the lemma to get $v = v + eh_2 = (v + h_1 + h_2) + h_2 = v + h_2$. ☐

Corollary 2.15. *The ghost surpassing relation on (supertropical) vector spaces is a partial order.*

Almost tangible vectors. Since one also has the example of supertropical algebras arising from tropicalizing Puiseux series [Izhakian and Rowen 2010], we digress briefly to discuss the situation when ν is not necessarily 1:1.

Remark 2.16. Define

$$\mathcal{T}_e := \{a \in \mathcal{T} : a \cong_\nu \mathbb{1}_R\}.$$

This is a submonoid of $\mathcal{T}$, and in fact $\mathcal{T}_e \cup \{e\}$ is a supertropical subsemifield† of F.

Lemma 2.17. *Generalizing Remark 2.6*(i), *an element* $a \in F$ *is tangible if and only if the following condition holds*: $a \models_{\mathrm{gs}} b$ *implies* $b = \alpha a$ *for some* $\alpha \in \mathcal{T}_e$.

Proof. Remark 2.6(i) yields the forward implication. Conversely, suppose a is not tangible, that is, $a \in \mathcal{G}$, so $a = a^\nu$. Then $a \models_{\text{gs}} \hat{a}$, where $\hat{a} \in \mathcal{T}$ and $(\hat{a})^\nu = a$. The condition implies $\hat{a} = \alpha a$ for some $\alpha \in \mathcal{T}_e$, which is impossible since $\alpha a \in \mathcal{G}$. $\square$

Motivated by Lemma 2.17, we have an abstract definition of tangibility for any (supertropical) vector space V over a supertropical semifield:

Definition 2.18. The *almost tangible vectors* of V are those elements $v \in V$ for which $v \models_{\text{gs}} w$, for all $w \in V$, implies $w \in \mathcal{T}_e v$.

Remark 2.19. A nonzero ghost vector v cannot be almost tangible. Indeed, we would have $v \models_{\text{gs}} \mathbb{0}$, implying $\mathbb{0} \in \mathcal{T}_e v$, a contradiction.

Example 2.20. Clearly, the almost tangible vectors of $F^{(n)}$ all are tangible.

On the other hand, in logarithmic notation, taking $F = D(\mathbb{R})$, if V is the submodule of $F^{(2)}$ spanned by the vectors $v_1 = (1, 1^\nu)$ and $v_2 = (0, 1)$, then one sees without difficulty that v_1 is almost tangible in V, although not tangible in $F^{(2)}$.

In fact, a subspace of $F^{(n)}$ need not have any tangible vectors at all, as exemplified by the submodule $F(1, 1^\nu)$ of $F^{(2)}$.

Here is a reduction to the case where ν is 1:1.

Remark 2.21. We define an equivalence on $F^\dagger$ via $a \equiv b$ when either $a = b$ or $a, b \in \mathcal{T}$ with $a \cong_\nu b$. In other words, two tangible elements are equivalent if and only if they are ν-matched. Then we could define the supertropical semifield† $F^\dagger/_\equiv$ to be $(\mathcal{T}/_\equiv) \cup \mathcal{G}$. The ghost map ν defines a 1:1 function from the equivalence classes of $\mathcal{T}$ to $\mathcal{G}$.

3. Background from matrices

Any set $S = \{v_1, \dots, v_m\}$ of m row vectors in $F^{(n)}$ corresponds to an $m \times n$ matrix $A(S)$, whose m rows are the vectors of S. We call $A(S)$ the *matrix* of S. We denote by $M_n(F)$ the monoid of $n \times n$ matrices over F, and by $M_n(\mathcal{G}_\mathbb{0})$ the monoid of $n \times n$ ghost matrices.

We recall that the *tropical determinant* of an $n \times n$ matrix $A = (a_{i,j})$ in $M_n(F)$ is really the permanent, which we denote as

$$|A| = \sum_{\pi \in S_n} a_{\pi(1),1} \dots a_{\pi(n),n}.$$

Although the equation $|AB| = |A||B|$ fails over the max-plus algebra, the relation $|AB| \models_{\text{gs}} |A||B|$ holds over a supertropical semiring by Theorem 3.5 of [Izhakian and Rowen 2011a], and any matrix satisfies its characteristic polynomial in the sense of Theorem 5.2 of the same work. The tangible roots of this polynomial are precisely the supertropical eigenvalues of A, as given in [Izhakian and Rowen 2011a, Theorem 7.10].

We say that the matrix A is *nonsingular* if $|A|$ is tangible (and thus A is quasi-invertible [Izhakian and Rowen 2011a]); otherwise, $|A| \in \mathcal{G}_{\mathbb{0}}$ (that is, $|A| \models_{\mathrm{gs}} \mathbb{0}_F$ by Remark 2.5) and we say that A is *singular*. Although it was shown in [Izhakian and Rowen 2011a] that the product of nonsingular matrices could be singular, we do have the consolation that the product of nonsingular matrices cannot be ghost; see Theorem 3.5 below.

In [Izhakian and Rowen 2011a], we also defined vectors in $F^{(n)}$ to be *tropically independent* if no linear combination with tangible coefficients is in $\mathcal{H}_{\mathbb{0}}$.

Recall that a *quasi-identity* matrix is a nonsingular, multiplicatively idempotent matrix ghost-surpassing the identity matrix; therefore its determinant equals $\mathbb{1}_F$. Suppose $A = (a_{i,j})$, with $|A|$ tangible. In [Izhakian and Rowen 2011b, Theorem 2.8] one defines the matrix

$$A^{\nabla} := \frac{\mathbb{1}_F}{|A|}\,\mathrm{adj}(A) \tag{3-1}$$

and obtains the quasi-identity matrices

$$I_A = AA^{\nabla}, \qquad I'_A = A^{\nabla}A. \tag{3-2}$$

Remark 3.1. Recall some results on nonsingularity and supertropical dependence:

- $A(S)$ has m tropically independent rows (resp. columns) if and only if $A(S)$ has a nonsingular $m \times m$ submatrix [Izhakian and Rowen 2009, Theorem 3.4; 2011a, Corollary 6.6]
- $|A\,\mathrm{adj}(A)| = |A|^n$ [Izhakian and Rowen 2011b, Theorem 4.9].
- $|\mathrm{adj}(A)| = |A|^{n-1}$ [Izhakian and Rowen 2011a]. Thus, $\mathrm{adj}(A)$ is nonsingular if A is nonsingular.

Thus, it is natural to try to understand linear algebra in terms of the supertropical matrix theory of [Izhakian and Rowen 2011a; 2011b].

Annihilators of matrices.

Definition 3.2. A vector $v \in V := F^{(n)}$ (written as a column) *g-annihilates* an $m \times n$ matrix A if $Av \models_{\mathrm{gs}} \mathbb{0}_V$ in V. Define

$$\mathrm{Ann}(A) = \{v \in V : Av \models_{\mathrm{gs}} \mathbb{0}_V\};$$

this is clearly a subspace of V.

Accordingly, $\mathcal{G}_{\mathbb{0}}^{(n)} \subseteq \mathrm{Ann}(A)$ for any $m \times n$ matrix A.

Remark 3.3. (i) The point of this definition is that the vector $v = (\beta_1, \dots, \beta_m)$ g-annihilates $(A(S))^{\mathrm{t}}$, the transpose of the matrix of $S = \{w_1, \dots, w_m\}$, if and only if $\sum_{i=1}^{m} \beta_i w_i \models_{\mathrm{gs}} \mathbb{0}_V$. Thus, tangible g-annihilators correspond to tropical dependence relations.

(ii) A (nonzero) tangible vector cannot g-annihilate a nonsingular matrix, since the columns are tropically independent.

We can improve this result to include vectors that are not necessarily tangible.

Lemma 3.4. *The diagonal of the product $I_A I_B$ of quasi-identity matrices I_A, I_B cannot all be ghosts.*

Proof. Recall from [Izhakian and Rowen 2011a, §3.2] the *weighted digraph* $G = (\mathscr{V}, \mathscr{E})$ of an $n \times n$ matrix $A = (a_{i,j})$, which is defined to have vertex set $\mathscr{V} = \{1, \dots, n\}$ and an edge (i, j) from i to j (of *weight* $a_{i,j}$) whenever $a_{i,j} \neq \mathbb{0}_F$.

Write $I_A = (a_{i,j})$ and $I_B = (b_{i,j})$. If the assertion is false, then for each i_t, the i_t, i_t diagonal entry must be a ghost, so there must be i_{t+1} such that $a_{i_t,i_{t+1}} b_{i_{t+1},i_t} \geq_\nu \mathbb{1}_F$. This means each edge from i_t to i_{t+1} in the weighted digraph G of $I_A I_B$ has weight $\geq_\nu 1$. By the pigeonhole principle, the path of vertices of $i_1, i_2, i_3, \dots, i_{n+1}$ contains a cycle, say from i_s to i'_s. But the weight of any nonloop cycle in a quasi-identity has ν-value $< \mathbb{1}_F$. (Otherwise, multiplying by the entries $a_{i,i}$ for all vertices i not in the cycle gives an extra summand $\geq e = \mathbb{1}_F^\nu$ for $|I_A|$, contrary to $|I_A| = \mathbb{1}_F$.) Hence

$$\mathbb{1}_F \leq_\nu \prod_{k=s}^{s'-1} a_{i_k,i_{k+1}} b_{i_{k+1},i_k} = \prod_{k=s}^{s'-1} a_{i_k,i_{k+1}} \prod_{k=s}^{s'-1} b_{i_{k+1},i_k} <_\nu \mathbb{1}_F \mathbb{1}_F = \mathbb{1}_F,$$

a contradiction. □

Theorem 3.5. *The product of two nonsingular $n \times n$ matrices cannot be in $M_n(\mathscr{G}_{\mathbb{0}})$.*

Proof. If AB is ghost for A, B nonsingular, then $(A^\nabla A)(BB^\nabla) \in M_n(\mathscr{G}_{\mathbb{0}})$, contradicting the lemma. □

On the other hand, examples were given in [Izhakian and Rowen 2011a] in which the product of two nonsingular $n \times n$ matrices is singular. Here is a related example using quasi-identities:

Example 3.6. The matrices

$$A = \begin{pmatrix} 0 & 0^\nu \\ -\infty & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & -\infty \\ 0^\nu & 0 \end{pmatrix}$$

over $D(\mathbb{R})$ are nonsingular, but $AB = \begin{pmatrix} 0^\nu & 0^\nu \\ 0^\nu & 0 \end{pmatrix}$ and $BA = \begin{pmatrix} 0 & 0^\nu \\ 0^\nu & 0^\nu \end{pmatrix}$ are singular.

4. Tropical dependence

Dependence plays a major role in module theory. The familiar definition becomes degenerate for supertropical vector spaces; the following modification from [Izhakian and Rowen 2011a], in which the role of zero is replaced by the ghost ideal, is more suitable for our purposes.

Definition 4.1. Suppose V is a supertropical vector space over F. A family of elements $S = \{w_i : i \in I\} \subset V$ is *tropically dependent* if there exists a nonempty finite subset $I' \subset I$ and a family $\{\alpha_i : i \in I'\} \subset \mathcal{T}$ such that

$$\sum_{i \in I'} \alpha_i w_i \in \mathcal{H}_{\mathbb{0}}. \tag{4-1}$$

Any such relation (4-1) is called a *tropical dependence* for S. A subset $S \subset V$ is called *tropically independent* if it is not tropically dependent.

Given an element $v \in V$, we say that v is *tropically dependent* on a family $S = \{w_i : i \in I\}$ if $S \cup \{v\}$ is tropically dependent, in which case we write $v \curlyvee_{\mathrm{gd}} S$. (In particular, $v \curlyvee_{\mathrm{gd}} \{v\}$.) A subset S' of V is *tropically dependent* on S if $v \curlyvee_{\mathrm{gd}} S$ for each $v \in S'$.

An easy observation:

Remark 4.2. Suppose $S = \{w_i : i \in I\} \subset V$. For any given set $\{\alpha_i : i \in I\} \subset \mathcal{T}$ of tangible elements of F, the set S is tropically independent if and only if $\{\alpha_i w_i : i \in I\}$ is tropically independent.

Lemma 4.3. *If a set S' is tropically dependent on $S \subset F^{(n)}$, then the supertropical vector space V of $F^{(n)}$ spanned by S' is also tropically dependent on S.*

Proof. It is enough to show that if $v, v' \in F^{(n)}$ are tropically dependent on S, then $v+v'$ is also tropically dependent on S. Write $v+\sum \alpha_i w_i \in \mathcal{H}$ and $v'+\sum \alpha_i' w_i \in \mathcal{H}$ for $w_i \in S$. It is enough to check this on each component. Thus, we may assume that v, v', w_i, and w_i' are all in F. If $v \cong_\nu v'$, then $v + v' \in \mathcal{G}$. On the other hand, if $v >_\nu v'$, then $v + v' = v$ is already tropically dependent on S, so the assertion is clear. □

Tropical d-bases and rank.

Definition 4.4. A *d-base* (for *dependence base*) of a supertropical vector space V is a maximal set of tropically independent elements of V. The *rank* of a d-base $\mathcal{B}$, denoted $\mathrm{rk}(\mathcal{B})$, is the number of elements of $\mathcal{B}$.

Our d-base corresponds to the "basis" in [Maclagan and Sturmfels 2009, Definition 5.2.4].

Proposition 4.5. *Any subspace of $F^{(n)}$ is tropically dependent on any subset S of n tropically independent elements. All d-bases of $F^{(n)}$ have precisely n elements.*

Proof. By Theorem 6.6 of [Izhakian and Rowen 2011a], the matrix A of S is nonsingular if and only if S is tropically independent, so in particular any d-base $\mathcal{B}$ of $F^{(n)}$ must have at least n elements. Also recall that any $n+1$ vectors of $F^{(n)}$ are tropically dependent, by Corollary 6.7 of the same work, so $\mathcal{B}$ has precisely n elements. □

This leads us to the following definition.

Definition 4.6. The *rank* of a supertropical vector space V is defined as

$$\operatorname{rk}(V) := \max\{\operatorname{rk}(\mathcal{B}) : \mathcal{B} \text{ is a d-base of } V\}.$$

For $V \subseteq F^{(n)}$, the *tangible rank* of V is defined as

$$\text{t-rk}(V) := \max\{\operatorname{rk}(\mathcal{B}) : \mathcal{B} \subset \mathcal{T}_{\mathbb{0}}^{(n)} \text{ is a d-base of } V\}.$$

We have just seen that $\text{t-rk}(F^{(n)}) = \operatorname{rk}(F^{(n)}) = n$.

Corollary 4.7. *If* $V \subset F^{(n)}$, *then* $\operatorname{rk}(V) \leq n$.

Proof. Any d-base of V is contained in a d-base of $F^{(n)}$ whose order must be that of the standard base given in (5-1) below, which is n. □

We might have liked $\operatorname{rk}(V)$ to be independent of the choice of d-base of V, for any supertropical vector space V. This is proved in the classical theory of vector spaces by showing that dependence is transitive. However, transitivity fails in the supertropical theory, since we have the following sort of counterexample.

Example 4.8. In logarithmic notation, over $D(\mathbb{R})^{(3)}$, the vector $v = (0, 1, 3)$ is tropically dependent on $W = \{w_1, w_2\}$, where $w_1 = (1, 1, 2)$ and $w_2 = (1, 1, 3)$, since $v + w_1 + w_2 = (1^\nu, 1^\nu, 3^\nu)$. Furthermore, W is tropically dependent on $U = \{u_1, u_2\}$, where $u_1 = (1, 1, 0)$ and $u_2 = (-\infty, -\infty, 1)$, since

$$w_1 + u_1 + 1u_2 = (1^\nu, 1^\nu, 2^\nu), \qquad w_2 + u_1 + 2u_2 = (1^\nu, 1^\nu, 3^\nu).$$

But v, u_1, and u_2 are tropically independent, since the tropical determinant of the matrix whose rows are these vectors is $3 \in \mathcal{T}$.

In fact, different d-bases may contain different numbers of elements, even when tangible. An example is given in [Maclagan and Sturmfels 2009, Example 5.4.20], which is reproduced here with different entries.

Example 4.9. Consider the following vectors in $D(\mathbb{R})^{(3)}$:

$$v_1 = (5, 5, 0), \quad v_2 = (5, 5, 4), \quad v_3 = (0, 1, 4), \quad v_4 = (0, 2, 4).$$

Then v_1, v_2, and v_3 are tropically dependent (since their sum $(5^\nu, 5^\nu, 4^\nu)$ is ghost), and likewise v_1, v_2, and v_4 are tropically dependent. It follows that $\{v_1, v_2\}$ is a d-base for the supertropical vector space V spanned by v_1, v_2, v_3, and v_4. But v_2, v_3, and v_4 are tropically independent since their determinant is 11, which is tangible; hence, $\{v_2, v_3, v_4\}$ is also a d-base of V.

We do have a consolation.

Lemma 4.10. *If the vectors $v_1, \dots, v_k \in F^{(n)}$ are tropically independent and the vector v is tangible, then there are $i_1, \dots, i_{k-1}$ in $\{1, \dots, k\}$ such that the vectors $v_{i_1}, \dots, v_{i_{k-1}}, v$ are tropically independent.*

Proof. Let A be the $(k+1) \times n$ matrix whose rows are $v_1, \dots, v_k, v$, and let A_0 denote the $k \times n$ matrix of the first k rows $v_1, \dots, v_k$. Then A_0 has a nonsingular $k \times k$ submatrix obtained by deleting $n-k$ columns; deleting these columns in A, we have reduced to the case that $n = k$, that is, A is a $(k+1) \times k$ matrix. Now let $A_0' = (a_{j,i}') = \operatorname{adj}(A_0)$, which is nonsingular; see Remark 3.1. We are done unless for each row $i \le k$, the $k \times k$ submatrix of A obtained by deleting the i row is singular, which means that $\sum_{j=1}^k a_{i,j}' a_{k+1,j}$ is ghost. This means that the vector $(a_{k+1,1}, \dots, a_{k+1,k})$ g-annihilates the nonsingular matrix A_0', which is impossible. □

Proposition 4.11. *For any tropical subspace V of $F^{(n)}$ and any tangible $v \in V$, there is a tangible d-base of V containing v whose rank is that of V.*

Proof. Take a tangible d-base of V of maximal rank, and apply the lemma. □

Example 4.12 (failure of the analog of Proposition 4.11 for nontangible vectors). Consider the supertropical vector space $W \subset D(\mathbb{R})^{(2)}$ spanned by $w_1 = (0, 1)$ and $w_2 = (0, 2)$. Then $v = (1, 3^\nu)$ comprises a d-base of W, consisting of only one element.

Proposition 4.13. *If A is a matrix of rank m, its g-annihilator has a tangible tropically independent set of rank $\ge n - m$.*

Proof. Take m tropically independent rows $v_1, \dots, v_m$ of A, which we may assume are the first m rows of A. For any other row v_u of A $(m < u \le n)$, we have $\beta_{u,1}, \dots, \beta_{u,m} \in \mathcal{T}_{\mathbb{0}}$ such that $v_u + \sum \beta_{i,j} v_i \in \mathcal{G}_{\mathbb{0}}^{(n)}$. Letting B be the $(n-m) \times n$ matrix whose (i, j) entries are $\beta_{i,j}$ for $1 \le i, j \le m$, and for which $\beta_{i,j} = \delta_{i,j}$ (the Kronecker delta) for $m < j \le n$, we see that B contains an $(n-m) \times (n-m)$ identity submatrix and so has tangible rank $\ge n - m$, but BA is ghost. □

Example 4.14. We exhibit a 3×3 matrix A over $D(\mathbb{R})$ of rank $m = 2$, all of whose entries are tangible, although $\operatorname{rk}(\operatorname{Ann}(A)) = 2 > 3 - 2$. Take

$$A = \begin{pmatrix} 4 & 4 & 0 \\ 4 & 4 & 1 \\ 4 & 4 & 2 \end{pmatrix}.$$

A is g-annihilated by the tropically independent vectors $v_1 = (1, 1, 0)^{\mathrm{t}}$ and $v_2 = (1, 1, 1)^{\mathrm{t}}$, since $Av_1 = Av_2 = (5, 5, 5)^{\mathrm{t}}$.

Note that this kind of example requires $n \ge 3$, in view of Theorem 3.5.

Saturated dependence relations. Let us study tropical dependence relations in $F^{(n)}$ more closely. Example 5.7(ii) below shows that a tropical dependence of a vector v on an independent set $S = \{w_i : i \in I\}$ is not uniquely determined. Nevertheless, in this subsection we do get a "canonical" tropical dependence relation, which we call *saturated*. But first, in order for tropical dependence relations to be well-defined with respect to the ghost map $\nu : F \to \mathcal{G}_{\mathbb{0}}$, we verify the following condition.

Lemma 4.15. *Any subspace of $F^{(n)}$ (with $\mathcal{H}_{\mathbb{0}} = \mathcal{G}_{\mathbb{0}}^{(n)}$) satisfies the property that whenever $\alpha_i, \beta_i \in \mathcal{T}$ with $\alpha_i \cong_\nu \beta_i$,*

$$\sum_i \alpha_i w_i \in \mathcal{H}_{\mathbb{0}} \iff \sum_i \beta_i w_i \in \mathcal{H}_{\mathbb{0}}. \tag{4-2}$$

Proof. The argument is analogous to that of Lemma 4.3. The condition clearly passes to submodules, so it is enough to prove it for $F^{(n)}$, and thus to check (4-2) on each component. We write $w_{i,j}$ for the j-component of w_i. Note that $\alpha_i w_{i,j} \cong_\nu \beta_i w_{i,j}$ for each i. There are two ways for $\sum_i \alpha_i w_{i,j} \in \mathcal{G}_{\mathbb{0}}$:

(1) Some $\alpha_{i'} w_{i',j}$ dominates $\sum_i \alpha_i w_{i,j}$ and is ghost, implying $w_{i',j} \in \mathcal{G}_{\mathbb{0}}$, so

$$\sum_i \beta_i w_{i,j} = \beta_{i'} w_{i',j} = \alpha_{i'} w_{i',j} \in \mathcal{G}_{\mathbb{0}}.$$

(2) Two essential summands $\alpha_{i'} w_{i',j}$ and $\alpha_{i''} w_{i'',j}$ are ν-matched. But then

$$\begin{aligned}\sum_i \beta_i w_{i,j} &= \beta_{i'} w_{i',j} + \beta_{i''} w_{i'',j} = (\beta_{i'} w_{i',j})^\nu \\ &= (\alpha_{i'} w_{i',j})^\nu = \alpha_{i'} w_{i',j} + \alpha_{i''} w_{i'',j} = \sum_i \alpha_i w_{i,j} \in \mathcal{G}_{\mathbb{0}}.\end{aligned}$$ □

We examine the tropical dependence

$$v \curlyvee_{\mathrm{gd}} \sum_{i \in I} \alpha_i w_i. \tag{4-3}$$

Lemma 4.16. *Suppose $V = F^{(n)}$. If $v \curlyvee_{\mathrm{gd}} \sum_{i \in I} \alpha_i w_i$ and $v \curlyvee_{\mathrm{gd}} \sum_{i \in I} \beta_i w_i$ for $\alpha_i, \beta_i \in \mathcal{T}_{\mathbb{0}}$, then taking $\gamma_i = \widehat{\alpha_i + \beta_i}$, we have*

$$v \curlyvee_{\mathrm{gd}} \sum_{i \in I} \gamma_i w_i.$$

Proof. Checking each component in turn, we may assume that $V = F$. We proceed as in Lemma 4.15. Namely, $v \curlyvee_{\mathrm{gd}} \sum_{i \in I} \alpha_i w_i$ (resp. $v \curlyvee_{\mathrm{gd}} \sum_{i \in I} \beta_i w_i$) implies one of the following:

(1) v and some term $\alpha_{i'} w_{i'}$ dominate (resp. v and $\beta_{i'} w_{i'}$ dominate), in which case $\gamma_{i'} = \alpha_{i'}$ (resp. $\gamma_{i'} = \beta_{i'}$).

(2) $\alpha_{i'}w_{i'}$ and $\alpha_{i''}w_{i''}$ dominate (resp. $\beta_{i'}w_{i'}$ and $\beta_{i''}w_{i''}$ dominate), in which case $\gamma_{i'} = \alpha_{i'}$ and $\gamma_{i''} = \alpha_{i''}$ (resp. $\gamma_{i'} = \beta_{i'}$ and $\gamma_{i''} = \beta_{i''}$).

(3) Some ghost term $\alpha_{i'}w_{i'}$ (resp. $\beta_{i'}w_{i'}$) dominates, in which case $\gamma_{i'} = \alpha_{i'}$ (resp. $\gamma_{i'} = \beta_{i'}$). □

Lemma 4.16 gives us a partial order on the coefficients of the tropical dependence relations of v on a set S, and motivates the following definition:

Definition 4.17. We say that the *support* of a tropical dependence

$$\alpha v \curlyvee_{\mathrm{gd}} \sum_{i \in I} \alpha_i w_i$$

(where $\alpha \in \mathcal{T}$ and $\alpha_i \in \mathcal{T}_{\mathbb{0}}$) is the set

$$\{i \in I : \alpha_i \neq \mathbb{0}_F\}.$$

A tropical dependence of minimal support is called *irredundant*.

A vector does not have a unique tropical dependence on a d-base. For example, if each entry of a vector w_1 is nonzero and $\alpha \in F$ has a suitably low ν-value, then $w_1 + \alpha w_2 = w_1$. We do have a slight consolation.

Lemma 4.18. *If $\sum_{i=1}^{k} \alpha_i w_i \curlyvee_{\mathrm{gd}} \sum_{i=1}^{k} \beta_i w_i$ for independent vectors $w_1, \dots, w_k$ and tangible α_i, β_i, then $\alpha_i \cong_\nu \beta_i$ for some i.*

Proof. $\sum_{i=1}^{k}(\alpha_i + \beta_i)w_i \in \mathcal{H}_{\mathbb{0}}$, implying that not all $\alpha_i + \beta_i$ are tangible, so that $\alpha_i \cong_\nu \beta_i$. □

We can do better. A tropical dependence of v on a tropically independent set S is called *saturated* if the coefficients α_i in (4-3) all are maximal possible with respect to $\geq_\nu$; in other words, whenever $v + \sum_{i=1}^{l} \beta_i w_i \in \mathcal{G}_{\mathbb{0}}^{(n)}$ with $\beta_i \in \mathcal{T}_{\mathbb{0}}$, then each $\beta_i \leq_\nu \alpha_i$.

Remark 4.19. If

$$v \curlyvee_{\mathrm{gd}} \sum_{i=1}^{l} \alpha_i w_i \tag{4-4}$$

is a saturated tropical dependence, then for any $k \leq l$ and for $v' = v + \sum_{i=1}^{k} \alpha_i w_i$,

$$v' \curlyvee_{\mathrm{gd}} \sum_{i=k+1}^{l} \alpha_i w_i \tag{4-5}$$

is also a saturated tropical dependence, since any ν-larger tropical dependence for (4-5) would yield the corresponding ν-larger tropical dependence for (4-4).

Theorem 4.20. *Any irredundant tropical dependence*

$$v \curlyvee_{\mathrm{gd}} \sum_{i=1}^{l} \alpha_i w_i \tag{4-6}$$

can be extended to a unique (up to equivalence in the sense of Remark 2.21) saturated tropical dependence of v on $S = \{w_1, \dots, w_l\}$, having the same support.

Remark 4.21. When the vector v is tangible and S is a d-base, Theorem 4.20 is an immediate consequence of [Izhakian and Rowen 2011b, Theorems 3.5 and 3.8], which shows that $Ax \models_{\mathrm{gs}} v$ has the maximal tangible vector solution $x = \widehat{\nu}(A^{\nabla} v)$, where $A^{\nabla} = (1/|A|)\,\mathrm{adj}(A)$. Here we take A to be the matrix of S, which is nonsingular, and x to be the vector $(\alpha_1, \dots, \alpha_l)^{\mathrm{t}}$.

In general, $x = A^{\nabla} v$ is a solution for the matrix equation $Ax \models_{\mathrm{gs}} v$, which, when v is written as a row, is $xA^{\mathrm{t}} \models_{\mathrm{gs}} v$. (In a sense, row form is more natural, since the matrix of S is obtained from the rows.) But this solution x need not be tangible.

Here is a direct combinatoric proof of Theorem 4.20 that does not rely on matrix theory, and does not depend on the additional assumption of tangibility of S.

Proof of Theorem 4.20. Uniqueness of a saturated tangible solution is obvious, since one could just take the sup of any two distinct saturated tropical dependences to get a contradiction. This also gives the motivation for proving existence. Write $v = (v_1, \dots, v_n)$. We start with some tropical dependence (4-6), which need not be saturated, with the aim of checking whether we can modify it until it is saturated. In principle, we could increase the ν-values of the coefficient α_i if at each component j of the vector $\alpha_i w_i$ the ν-value of v_j is not attained, and this is the main idea behind the proof. But increasing α_i still may not yield a saturated tropical dependence, since the coefficient may be allowed to increase further, so long as some other term in the tropical dependence also is adjusted so as to have a j-component of the same ν-value. Since these j-components are the most difficult to keep track of, we pay special attention to them. Write $w_{i,j}$ for the j-component of w_i.

We say that an index $j \le n$ has ***type 1*** if v_j is not dominated by $\sum_i \alpha_i w_{i,j}$, which means that either v_j itself is ghost, or else there is precisely one i with $\alpha_i w_{i,j}$ matching v_j, and this $w_{i,j} \in \mathcal{T}$.

We say that j has ***type 2*** for v if v_j is dominated by $\sum_i \alpha_i w_{i,j}$, which means that either there exists i such that $\alpha_i w_{i,j}$ is ghost and dominates v_j or there are i, i' such that $\alpha_i w_{i,j}$ and $\alpha_{i'} w_{i',j}$ are ν-matched and both dominate v_j.

Note that increasing the coefficients α_i in a tropical dependence cannot change the type of an index j from type 2 to type 1. Also, at least one index must have type 1, since otherwise $\sum \alpha_i w_{i,j} \in \mathcal{G}_{\mathbb{0}}^{(n)}$, contrary to the hypothesis that the w_i are tropically independent. We choose our tropical dependence such that the number of indices of type 1 is minimal. In this case, if $\alpha_i w_{i,j}$ ν-matches v_j for j of type 1, we cannot find a ν-greater tropical dependence in which α_i is increased, since this would force the tropical dependence to have an extra type 2 index. Thus, in this case we say w_i is *anchored* at j. Renumbering the vectors, we may assume that $w_1, \dots, w_k$ are anchored at various indices, and replace v by $v' = v + \sum_{i=1}^k \alpha_i w_i$. Now we have a new tropical dependence

$$v' + \sum_{i=k+1}^{l} \alpha_i w_i \in \mathcal{G}_{\mathbb{0}}^{(n)},$$

which by induction on l can be extended to a saturated tropical dependence

$$v' \curlyvee_{\mathrm{gd}} \sum_{i=k+1}^{l} \alpha_i' w_i.$$

But then the tropical dependence

$$v \curlyvee_{\mathrm{gd}} \left(\sum_{i=1}^{k} \alpha_i w_i + \sum_{i=k+1}^{l} \alpha_i' w_i \right)$$

is saturated, since $w_1, \dots, w_k$ are anchored. □

Proposition 4.22. *If*

$$v \curlyvee_{\mathrm{gd}} \sum_{i=1}^{l} \alpha_i w_i \quad \textit{and} \quad v' \curlyvee_{\mathrm{gd}} \sum_{i=1}^{l} \alpha_i' w_i \tag{4-7}$$

are saturated tropical dependences, then

$$(v + v') \curlyvee_{\mathrm{gd}} \sum_{i=1}^{l} (\widehat{\alpha_i + \alpha_i'}) w_i \tag{4-8}$$

is also a saturated tropical dependence.

Proof. Again we have two proofs, the first using results from [Izhakian and Rowen 2011b] in the case when v, v' are tangible and the matrix A of the w_i is nonsingular. In the first case, one just takes the solutions $x = \widehat{A^\nabla v}$ and $x' = \widehat{A^\nabla v'}$ for the vectors accompanying the α_i and the α_i', and then notes that

$$\widehat{\nu}\big(A^\nabla v + A^\nabla v'\big) = \widehat{\nu}\big(A^\nabla (v + v')\big).$$

For the general case, one needs to modify the second proof of Theorem 4.20 for the vector $v + v'$. Namely, consider the tropical dependence

$$(v + v') \curlyvee_{\mathrm{gd}} \sum_{i=1}^{l} \gamma_i w_i,$$

where $\gamma_i = (\widehat{\alpha_i + \alpha_i'})$. At least one index in this tropical dependence must have type 1 for $v+v'$, since otherwise the w_i are tropically dependent. We choose our tropical dependence such that the number of indices of type 1 is minimal. As before, if $\gamma_i w_{i,j}$ ν-matches v_j for j of type 1, we cannot find a larger tropical dependence in which γ_i is increased, so w_i is anchored at j. Again, we may assume that $w_1, \dots, w_k$ are anchored at various indices, and replace $v+v'$ by $v'' = v+v'+\sum_{i=1}^{k} \gamma_i w_i$. But

$$\left(v + \sum_{i=1}^{k} \alpha_i w_i\right) \curlyvee_{\mathrm{gd}} \sum_{i=k+1}^{l} \alpha_i w_i \quad \text{and} \quad \left(v' + \sum_{i=1}^{k} \alpha'_i w_i\right) \curlyvee_{\mathrm{gd}} \sum_{i=k+1}^{l} \alpha'_i w_i$$

are saturated tropical dependences by Remark 4.19, so by induction on l,

$$v'' \curlyvee_{\mathrm{gd}} \sum_{i=k+1}^{l} \gamma_i w_i$$

is a saturated tropical dependence. But then the tropical dependence

$$v \curlyvee_{\mathrm{gd}} \left(\sum_{i=1}^{k} \gamma_i w_i + \sum_{i=k+1}^{l} \gamma_i w_i\right)$$

is saturated. □

5. Tropical spanning

In this section, we consider further the fundamental question of what “base” should mean for supertropical vector spaces. The d-base (defined above) competes with another notion to be obtained from $\models_{\mathrm{gs}}$. But for the moment we turn to the naive analog from the classical theory of linear algebra.

Definition 5.1 (classical bases and the standard base). A (supertropical) vector space V over a semifield F is *classically spanned* by a set $S = \{w_i : i \in I\}$ if every element of V can be written in the form

$$v = \sum_{i \in J} \alpha_i w_i,$$

for $\alpha_i \in F$ and some finite index set $J \subset I$.

A set $\mathcal{B} = \{b_1, \dots, b_n\} \subset V$ is a *classical base* of a vector space V over a semifield F if every element of V can be written uniquely in the form $\sum_{i=1}^{n} \alpha_i b_i$ for $\alpha_i \in F$.

The *standard base* of $F^{(n)}$ is the classical base defined as

$$\begin{aligned} \varepsilon_1 &= (\mathbb{1}_F, \mathbb{0}_F, \dots, \mathbb{0}_F), \\ \varepsilon_2 &= (\mathbb{0}_F, \mathbb{1}_F, \mathbb{0}_F, \dots, \mathbb{0}_F), \\ &\dots \\ \varepsilon_n &= (\mathbb{0}_F, \mathbb{0}_F, \dots, \mathbb{1}_F). \end{aligned} \tag{5-1}$$

Proposition 5.2. *If V has a classical base $b_1, \dots, b_n$, then V is isomorphic to $F^{(n)}$.*

The proof is standard; one defines the isomorphism $F^{(n)} \to V$ by

$$(\alpha_1, \dots, \alpha_n) \mapsto \sum_{j=1}^{n} \alpha_j b_j.$$

Definition 5.3 (tropical spanning). A vector $v \in V$ is *tropically spanned* by a set $S = \{w_i : i \in I\} \subset V$ if there exist a nonempty finite subset $J \subset I$ and a family $\{\alpha_i : i \in J\} \subset \mathcal{T}$ such that

$$v \models_{\mathrm{gs}} \sum_{i \in J} \alpha_i w_i. \tag{5-2}$$

In this case, we write $v \models_{\mathrm{gs}} S$.

A subset $S' \subseteq V$ is *tropically spanned* by S, written $S' \models_{\mathrm{gs}} S$, if $v \models_{\mathrm{gs}} S$ for each $v \in S'$.

Remark 5.4 (transitivity for tropical spanning). If $V \models_{\mathrm{gs}} W$ and $W \models_{\mathrm{gs}} U$, then $V \models_{\mathrm{gs}} U$.

Obviously, any set classically spanned by S is tropically spanned; surprisingly, the converse often holds.

Remark 5.5.
(i) If a tangible vector $v \in F^{(n)}$ is tropically spanned by a set $S \subset V$, then v is classically spanned by S with the same coefficients, as seen by checking components.

(ii) The assertion in (i) can fail for nontangible $v \in F^{(2)}$. Take $S = \{(\mathbb{1}_F, \mathbb{1}_F)\}$; then $v = (\mathbb{1}_F, \mathbb{1}_F^{\nu})$ is tropically spanned by S since $(\mathbb{1}_F, \mathbb{1}_F^{\nu}) \models_{\mathrm{gs}} (\mathbb{1}_F, \mathbb{1}_F^{\nu})$, but is not classically spanned by S.

(iii) If V has a classical spanning set $\mathcal{B}$ of almost tangible vectors, and $\mathcal{B}$ is tropically spanned by a set S, then V is classically spanned by S, by (ii) and transitivity. In particular, if $F^{(n)}$ is tropically spanned by a set S, then $F^{(n)}$ is classically spanned by S, since $F^{(n)}$ has the standard base.

(iv) Any element tropically spanned by S is also tropically dependent on S, but not conversely; for example $v = (\mathbb{1}_F, \mathbb{1}_F) \in F^{(2)}$ is tropically dependent on $S = \{(\mathbb{1}_F, \mathbb{1}_F^{\nu})\} \subset F^{(2)}$, but v is not tropically spanned by S. This leads to an interesting dichotomy to be studied shortly.

(v) Tropical spanning does not satisfy the assertion analogous to Lemma 4.16. For example, take

$$\big\{w_1 = (1, 2), w_2 = (1, 3)\big\} \subset D(\mathbb{R})^{(2)}$$

and the vector $v = (1, 3^{\nu})$; then $v \models_{\mathrm{gs}} w_1$ and $v \models_{\mathrm{gs}} w_2$, but $v \not\models_{\mathrm{gs}} w_1 + w_2 = (1^{\nu}, 3)$.

Thus, we see that almost tangible vectors already begin to play a special role in the theory of tropical dependence, and could be used instead of tangible vectors in the general theory of supertropical vector spaces.

Lemma 5.6. *$W = \{v \in V : v \models_{\mathrm{gs}} S\}$ is a subspace of V @for any $S \subset V$.*

Proof. If $v = \sum_{i \in I} \alpha_i w_i + y$ and $v' = \sum_{i \in I} \alpha_i' w_i + z$, where $\alpha_i, \alpha_i' \in \mathcal{T}$, $w_i \in S$ and $y, z \in \mathcal{H}_{\mathbb{0}}$, then letting $J = \{i : \alpha_i \cong_\nu \alpha_i'\}$, we have, by bipotence,

$$v + v' = \sum_{i \notin J} \beta_i w_i + \sum_{i \in J} \alpha_i^{\nu} w_i + (y + z) \models_{\mathrm{gs}} \sum_{i \notin J} \beta_i w_i,$$

where $\beta_i \in \{\alpha_i, \alpha_i'\} \subset \mathcal{T}$. The other verifications are easier. $\square$

We call W (in Lemma 5.6) the subspace *tropically spanned* by S, and say that S is a *tropically spanning set* of W.

A supertropical vector space is *finitely spanned* if it has a finite tropically spanning set.

Example 5.7. Take $R = D(\mathbb{R})$, with logarithmic notation.

(i) The vectors

$$v_1 = (1, 0, 1), \quad v_2 = (1, 1, 0), \quad \text{and} \quad v_3 = (0, 1, 1)$$

are tropically dependent in $D(\mathbb{R})^{(3)}$ since their sum is $(1^\nu, 1^\nu, 1^\nu)$. None of these vectors is tropically spanned by the two other vectors.

(ii) Even when a vector is classically spanned by tropically independent vectors, the coefficients need not be unique. For example,

$$(4, 5) = 2(1, 1) + 2(2, 3) = 1(1, 1) + 2(2, 3).$$

The point of this example is that the first coefficient is sufficiently small so as not to affect the outcome.

(iii) Another such example: The vectors

$$v_1 = (-\infty, -\infty, 1), \quad v_2 = (1, 1, -\infty), \quad \text{and} \quad v_3 = (-\infty, 1, 1)$$

are tropically independent, though classical spanning with respect to them (and thus also tropical spanning) is not unique; e.g., $(3, 3, 1) = 2v_2 + v_3 = v_1 + 2v_2$.

(iv) Another such example: Consider the vectors

$$v_1 = (1, 4, 3), \quad v_2 = (2, 3, 4), \quad \text{and} \quad v_3 = (0, 20, 20).$$

Then $(3, 20, 20) = 1v_2 + v_3 = 2v_1 + v_3$.

For $S = \{w_1, \dots, w_n\}$, there need not be a ν-maximal set of $\alpha_1, \dots, \alpha_l \in \mathscr{T}$ such that $v \models_{\mathrm{gs}} \sum_{i=1}^{l} \alpha_i w_i$. For example, in logarithmic notation, take

$$v = (1, 1), \quad w_1 = (1, 0), \quad \text{and} \quad w_2 = (1, 1).$$

Then $v = \alpha w_1 + w_2$ for all $\alpha < 0$, but taking $\alpha = 0$ yields $w_1 + w_2 = (1^\nu, 1)$.

Proposition 5.8. *For any subspace V of $F^{(n)}$, the number of elements of any tropically spanning set S of V is at least* $\mathrm{rk}(V)$.

Proof. Take a d-base $\{v_1, \dots, v_m\}$ of V, where $m = \mathrm{rk}(V) \le n$. By [Izhakian and Rowen 2009, Theorem 3.4], the $m \times n$ matrix whose rows are $v_1, \dots, v_m$ has rank m. Taking a nonsingular $m \times m$ submatrix and erasing all the $n - m$ columns not appearing in this submatrix, we may assume that $m = n$ (since we still have a supertropically generating set which we can shrink to a minimal one).

Writing $v_i \models_{\mathrm{gs}} \sum \alpha_{i,j} s_j$ for suitable $s_j \in S$, we can procure a nonsingular matrix whose rows are various s_j, implying that some subset of m vectors of S is tropically independent, and thus $|S| \ge m$. ☐

We are ready for another version of base.

Definition 5.9 (s-base). An *s-base* (for *spanning base*) of a (supertropical) vector space V is a minimal tropical spanning set S, in the sense that no proper subset of S tropically spans V.

As we shall see in Examples 5.22 below, a vector space with a finite d-base could still fail to have an s-base. Even when an s-base exists, it could be considerably larger than any d-base.

Example 5.10. Elements of a vector space V may be tropically dependent on a subspace W but not tropically spanned by W, as indicated in Example 5.7(i).

Example 5.11. Let V be the subspace of $D(\mathbb{R})^{(2)}$ spanned by

$$S = \{(1, 1), (1^\nu, 1), (1, 1^\nu)\}$$

in logarithmic notation, equipped with the standard ghost module. Each of these vectors alone comprises a d-base of V, whereas S is an s-base of V.

An s-base S need not be finite. On the other hand, obviously any finite tropical spanning set contains an s-base, so any finitely spanned vector space has an s-base. In order to coordinate the definitions of s-base and d-base, we introduce the following definition.

Definition 5.12. A *d,s-base* is an s-base which is also independent, that is, also is a d-base. A supertropical vector space V is *finite-dimensional* if it has a finite d,s-base.

Proposition 5.13. *The cardinality of the d,s-base S, if finite, is precisely* $\mathrm{rk}(V)$.

Proof. $|S| \geq \mathrm{rk}(V)$ by Proposition 5.8. But we get equality, since by definition S is itself a d-base. □

Example 5.14. Suppose S is a tropically independent subset of V. Then S is a d,s-base of the subspace of V tropically spanned by S. These are the subspaces of greatest interest to us.

Example 5.15. There are four possible sorts of nonzero subspaces of $F^{(2)}$ tropically spanned by a set S of tangible elements over a supertropical semifield F, writing $\{\varepsilon_1 = (\mathbb{1}_F, \mathbb{0}_F),\ \varepsilon_2 = (\mathbb{0}_F, \mathbb{1}_F)\}$ for the standard base:

(i) The plane $F^{(2)}$ itself.

(ii) A half-plane of tangible rank 2, having tangible s-base containing ε_1 or ε_2, as well as one tangible element $\alpha_1\varepsilon_1 + \alpha_2\varepsilon_2$ for $\alpha_1, \alpha_2 \in \mathcal{T}$.

(iii) A planar strip of tangible rank 2, having tangible s-base

$$\{\alpha_1\varepsilon_1 + \alpha_2\varepsilon_2,\ \beta_1\varepsilon_1 + \beta_2\varepsilon_2\},$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathcal{T}$.

(iv) A subspace of tangible rank 1, each pair of whose elements is tropically dependent. The tangible vectors are all multiples of a single vector.

One also has examples of nontangibly generated subspaces of $F^{(2)}$, such as $W = \{(\alpha, \alpha^\nu) : \alpha \in F\}$.

Here are some examples of strange behavior of s-bases, when we reverse the direction in tropical spanning in a supertropical vector space V. For subsets $S, S' \subset V$, we say that S' *ghost surpasses* S if for each $v \in S'$, there is some $v' \in S'$ such that $v' \models_{\mathrm{gs}} v$.

Examples 5.16. (i) Let V be the vector space tropically spanned by the vectors v_1, v_2, v_3 of Example 5.7(i), and let W be the subspace tropically spanned by the vectors v_1 and v_2. Note that

$$v_1 + v_2 = (1^\nu, 1, 1) \models_{\mathrm{gs}} (0, 1, 1) = v_3.$$

We claim more generally that W ghost surpasses V. Indeed, by symmetry we may consider $v = (\alpha, \beta, \gamma) \in V$ with $\alpha \leq_\nu \beta \leq_\nu \gamma$. By the definition of V, we must have either $\beta \cong_\nu \gamma$ or $\gamma = \gamma^\nu$. In the former case, $\beta(v_1 + v_2) \models_{\mathrm{gs}} v$, and in the latter case, $v \models_{\mathrm{gs}} \beta v_1 + \gamma v_3$.

(ii) Here is an example of a supertropical space V with an s-base $\{v_1, v_2, v_3, v\}$ in which the subspace W tropically spanned by v_1, v_2, and v_3 ghost surpasses $\{v\}$, and the subspace W' tropically spanned by v_1 and v_2 ghost surpasses $\{v_3\}$, but W' does not ghostsurpass $\{v\}$. Let

$$v_1 = (2, 2, 0, 2, 2, 0), \quad v_2 = (2, 0, 2, 2, 0, 2),$$
$$v_3 = (0, 2, 2, 0, 2, 2), \quad v = (2, 2, 2, 1, 1, 1).$$

Then $v_1 + v_2 + v_3 \models_{\mathrm{gs}} v$ and $v_1 + v_2 \models_{\mathrm{gs}} v_3$, but we claim that W' does not ghost surpass $\{v\}$. Suppose $\alpha_1 v_1 + \alpha_2 v_2 \models_{\mathrm{gs}} v$. If $\alpha_1 = \alpha_2$, then from the second component, $\alpha_1 = 0$, so from the fifth component $\alpha_2 = 2$: a contradiction.

Thus $\alpha_1 \neq \alpha_2$. By symmetry, we may assume that $\alpha_1 > \alpha_2$. Then from the first component, $\alpha_1 = -2$, and now we have a contradiction from the second component.

(iii) We now exhibit a space V in which the s-base $\mathcal{B}$ has 6 vectors and which is ghost surpassed by a set of 4 vectors; but if one removes any single vector from $\mathcal{B}$ it fails to ghost surpass V. Take

$$v_1 = (4, 4, 0, 2, 2, 0), \quad v_2 = (4, 0, 4, 2, 0, 2), \quad v_3 = (0, 4, 4, 0, 2, 2),$$
$$w_1 = (2, 2, 0, 2, 2, 0), \quad w_2 = (2, 0, 2, 2, 0, 2), \quad w_3 = (0, 2, 2, 0, 2, 2).$$

The vector $(4, 4, 4, 3, 3, 3)$ is not ghost surpassed by any 5 of these vectors.

Critical elements versus s-bases. Since s-bases are involved in the actual generation of the space, they are more in tune with the classical theory of convexity, and can be studied combinatorially. Here is another way to view the s-base, which is inspired by the literature on convex spaces. We say that two elements v, w in a supertropical vector space V over a supertropical semifield F are *projectively equivalent*, written $v \sim w$, if and only if $v = \alpha w$ for some tangible element $\alpha \in F$. Accordingly, we define the equivalence class of v as

$$[v]_\sim := \{w \in V \mid w \sim v\}.$$

Definition 5.17. A vector $v \notin \mathcal{H}_0$ in a supertropical vector space V is *critical* if we cannot write $v \models_{\mathrm{gs}} v_1 + v_2$ for $v_1, v_2 \in V \setminus [v]_\sim$. Taking one representative for each class $[v]_\sim$, a *critical set* of V is defined as a set of representatives of all the critical elements of V.

Critical elements correspond to "extreme points" over the max-plus algebra in [Gaubert and Katz 2007], in which it is shown that every point in $F^{(n)}$ is a linear combination of at most $n+1$ extreme points.

Example 5.18. Consider the space V spanned by the five critical vectors

$$(0, -\infty, 0, -\infty, 0, -\infty), \quad (-\infty, 0, -\infty, 0, -\infty, 0),$$
$$(0, -\infty, -\infty, 0, -\infty, -\infty), \quad (-\infty, 0, -\infty, -\infty, 0, -\infty),$$
$$(-\infty, -\infty, 0, -\infty, -\infty, 0).$$

Then (0, 0, 0, 0, 0, 0) is the sum of the first two vectors as well as the last three.

There is a basic connection between criticality and almost tangible.

Lemma 5.19. *Suppose $v \models_{\mathrm{gs}} \alpha v + w$ for $\alpha \in \mathcal{T}$, $v, w \in V$. Then $\alpha \le_\nu e$. Moreover:*

(1) *If $\alpha <_\nu e$, then $v \models_{\mathrm{gs}} w$.*

(2) *Suppose $\alpha \in \mathcal{T}_e$, that is, $\alpha \cong_\nu e$. If $w \in \mathcal{H}_{\mathbb{0}}$, then $v = \alpha v$. For any $w \in V$,*

$$v = \alpha^2 v + e w' = \alpha^2 v,$$

where $w' \models_{\mathrm{gs}} w$.

Proof. Write $v = \alpha v + w'$, where $w' \models_{\mathrm{gs}} w$.

If $\alpha >_\nu e$, then

$$v = \alpha v + w' = (\alpha + \mathbb{1}_F)v + w' = v + \alpha v + w' = v + v = ev \in \mathcal{H}_{\mathbb{0}}.$$

But then $v = v + \alpha v + w'$, implying by Proposition 2.14 that $v = v + \alpha v = \alpha v$, and thus $\alpha \cong_\nu e$.

(1) If $\alpha <_\nu e$, then $\mathbb{1}_F = \alpha + \mathbb{1}_F$, implying

$$v = (\alpha + \mathbb{1}_F)v = \alpha v + v = \alpha v + \alpha v + w' = e\alpha v + w' \models_{\mathrm{gs}} w,$$

proving (1).

(2) Thus, we assume that $\alpha \in \mathcal{T}_e$. If $w = ew$, then

$$v = \alpha v + w' = \alpha(\alpha v + w') + w' = \alpha^2 v + (\alpha + \mathbb{1}_F)w' = \alpha^2 v + ew' = \alpha(\alpha v + ew') = \alpha v.$$

For any w, if $\alpha \in \mathcal{T}_e$, then

$$v = \alpha v + w' = \alpha(\alpha v + w') + w' = \alpha^2 v + (\alpha + \mathbb{1}_F)w' = \alpha^2 v + ew'.$$

Hence, $v = \alpha^2 v$ by the preceding argument, replacing w by ew'. □

Proposition 5.20. *Any critical element $v \in V$ is almost tangible.*

Proof. Otherwise $v = w + w'$ for suitable $w \in V$, $w' \in \mathcal{H}_{\mathbb{0}}$, for which $w \notin \mathcal{T}_e v$, but by criticality, $w = \alpha v$ for $\alpha \in \mathcal{T}$. Then, by Lemma 5.19, $\alpha \le_\nu e$, and furthermore $\alpha \in \mathcal{T}_e$, since otherwise $v \models_{\mathrm{gs}} ew'$, contrary to $v \notin \mathcal{H}_{\mathbb{0}}$. But now, by Lemma 5.19, $v = \alpha v = w$, a contradiction. □

Lemma 5.21. *An almost tangible element $v \in V$ is critical if and only if it is not tropically spanned by $V \setminus [v]_\sim$, that is, $v \not\models_{\mathrm{gs}} \sum \alpha_i w_i$ for any $\alpha_i \in \mathcal{T}$, $w_i \in V \setminus [v]_\sim$.*

Proof. The "if" part is immediate by definition. To prove the converse, suppose on the contrary that $v \models_{\mathrm{gs}} \sum_{i=1}^t \alpha_i w_i$; by definition of criticality, $t > 1$. Then taking $v_1 = \alpha_1 w_1$ and $v_2 = \sum_{i=2}^t \alpha_i w_i$, we must have $v_2 \in [v]_\sim$, and conclude by induction on t. □

Clearly a critical set of a vector space V is projectively unique, but could be empty.

Examples 5.22. (i) The standard base $\{\varepsilon_1, \dots, \varepsilon_n\}$ of $F^{(n)}$ is also its critical set.

(ii) The critical set of the subspace $W = F^{(2)} \setminus ([\varepsilon_1]_\sim \cup [\varepsilon_2]_\sim)$ is empty.

(iii) $W = F^{(2)} \setminus [\varepsilon_1]_\sim$ has the critical set $[\varepsilon_2]_\sim$, but has no s-base.

Despite the last two examples, some positive information is available.

Lemma 5.23. *Any tropical spanning set S contains a tropical critical set of V.*

Proof. Suppose $v \in V$ is critical. By hypothesis on S, v is tropically spanned by S, but by Lemma 5.21, it must be an element of S (up to projective equivalence). □

Theorem 5.24. *Suppose V has an s-base S. Then S is a critical set of V.*

Proof. In view of Lemma 5.23, it remains to show that each element of S is critical. Suppose $v \in S$ is not critical. Then $v = v_1 + v_2$ where $v_1, v_2 \notin \mathcal{T} v$. Thus, when we write

$$v_1 = \sum_i \alpha_{1,i} s_{1,i} + w_1 \quad \text{and} \quad v_2 = \sum_i \alpha_{2,i} s_{2,i} + w_2$$

for $\alpha_{1,i}, \alpha_{2,i} \in \mathcal{T}$, $s_{1,i}, s_{2,i} \in S$, and $w_1, w_2 \in \mathcal{H}_{\mathbb{0}}$, we must have v appearing in one of the sums (for otherwise $v = v_1 + v_2$ is tropically spanned by the other elements of S, contrary to hypothesis).

Thus, we may assume $s_{1,1} = v$, and we have

$$v_1 \models_{\text{gs}} \alpha_1 v + \sum_{i \neq 1} \alpha_{1,i} s_{1,i},$$

and similarly $v_2 \models_{\text{gs}} \alpha_2 v + \sum_{i \neq 1} \alpha_{2,i} s_{2,i}$. (Formally, we permit $\alpha_2 = \mathbb{0}_F$.) We also write $v_j = \alpha_j + x_j$, where $x_j \models_{\text{gs}} \sum_{i \neq 1} \alpha_{j,i} s_{j,i}$.

Now

$$v = v_1 + v_2 = \beta v + x,$$

where $\beta = \alpha_{1,1} + \alpha_{2,1}$ and $x = x_1 + x_2$. But then $\beta \leq_\nu e$, by Lemma 5.19, which also says that if $\beta <_\nu e$, then $v \models_{\text{gs}} x$, contrary to S being an s-base. Thus, we may conclude that $\beta \cong_\nu e$. By symmetry, we assume that $\alpha_1 \cong_\nu e$. If $\alpha_2 <_\nu e$, then $v_2 \models_{\text{gs}} x_2$, and

$$\begin{aligned} v = v_1 + v_2 = \alpha_1 v + x_1 + v_2 &= \alpha_1(\alpha_1 v + x_1 + v_2) + x_1 + v_2 \\ &= \alpha_1^2 v + (\alpha_1 + \mathbb{1}_F)(x_1 + v_2) = \alpha_1^2 v + e(x_1 + v_2), \end{aligned}$$

and thus

$$\begin{aligned} v_1 = \alpha_1^2 v + e(x_1 + v_2) &= \alpha_1(\alpha_1^2 v + e(x_1 + v_2)) + x_1 + v_2 \\ &= \alpha_1^3 v + e(x_1 + v_2) = \alpha_1(\alpha_1^2 v + e(x_1 + v_2)) = \alpha_1 v. \end{aligned}$$

Thus, we are done for $\alpha_2 <_\nu e$, and may assume that $\alpha_2 \in \mathcal{T}_e$. Then

$$v = (\alpha_1 + \alpha_2)v + x_1 + x_2 = ev + x,$$

implying $v = ex \in \mathcal{H}$ and thus $\alpha_j v = \alpha_j ev = ev$ for $j = 1, 2$. Hence

$$v = v_1 + v_2 = \alpha_1 v + x_1 + \alpha_2 v + x_2 = (\alpha_1 + \alpha_2)v + x = ev + x,$$

and thus $v = ev + ex$ by Lemma 5.19, implying

$$\begin{aligned} v_1 &= ev + x_1 = ev + ex + x_1 = ev + ex_1 + ex_2 + x_1 \\ &= ev + ex_1 + ex_2 = ev + ex = v. \end{aligned}$$ □

Thus, we have the following striking result:

Corollary 5.25. *The s-base (if it exists) of a supertropical vector space is unique up to multiplication by tangible elements of F, and is comprised of almost tangible elements.*

Example 5.26. The only s-bases of the supertropical vector space $V = F^{(n)}$ are its classical bases $S = \{\alpha_1\varepsilon_1, \dots, \alpha_n\varepsilon_n\}$, where $\alpha_1, \dots, \alpha_n \in \mathcal{T}$.

One also has the following tie between critical sets and s-bases.

Proposition 5.27. *Any critical set C of a supertropical vector space V is an s-base of the subspace W tropically spanned by C.*

Proof. By hypothesis, C tropically spans W, so we need only check minimality. But for any $v \in C$, by definition, $C \setminus \{v\}$ does not tropically span v. □

Definition 5.28 (thick subspace). A subspace W of a supertropical vector space V is *thick* if $\mathrm{rk}(W) = \mathrm{rk}(V)$.

For example, any subspace of $F^{(n)}$ containing n tropically independent vectors is thick.

Remark 5.29. By definition, any thick subspace of a thick subspace of V is thick in V.

Remark 5.30. Any thick subspace W of a supertropical vector space V contains a d-base of V. Indeed, by definition, for $n = \mathrm{rk}(V)$, W contains a set of n tropically independent elements, which must be a maximal tropically independent set in V, by definition of rank. Thus, V is tropically dependent on any thick subspace.

Example 5.31. There exists an infinite chain of thick subspaces $W_1 \subset W_2 \subset \cdots$ of $V = D(\mathbb{R})^{(2)}$, where W_k is the strip tropically spanned by $\{(k, 0), (0, k)\}$, $k \in \mathbb{N}^+$. Thus, $\{(k, 0), (0, k)\}$ is not an s-base of $D(\mathbb{R})^{(2)}$. (One could expand this to an uncountable chain by taking $k \in \mathbb{R}^+$.)

Note that projectively a thick subspace need not be either convex in the classical sense or of pure dimension [Izhakian et al. 2011].

5A. ***Change of base matrices.*** We write P_π for the permutation matrix whose entry in the $(i, \pi(i))$ position is $\mathbb{1}_F$ (for each $1 \le i \le n$) and $\mathbb{0}_F$ elsewhere. Likewise, we write $\operatorname{diag}\{a_1, \dots, a_n\}$ for the diagonal matrix whose entry in the (i, i) position is a_i and $\mathbb{0}_F$ elsewhere, and denote it as D. We call the product $P_\pi D$ of a permutation matrix and a tangible (nonsingular) diagonal matrix, with each diagonal entry $\neq \mathbb{0}_F$, a *generalized permutation* matrix, and denote it as $\widetilde{P}_{\pi;D}$.

Recall from [Izhakian and Rowen 2011a, Proposition 3.9] that over a supertropical semifield, a matrix is invertible if and only if it is a generalized permutation matrix $\widetilde{P}_{\pi;D}$ with D nonsingular. In particular, the set of all generalized permutation matrices form a group whose unit element is I.

Definition 5.32. Given s-bases $\mathcal{B} = \{v_1, \dots, v_n\}$ and $\mathcal{B}' = \{v'_1, \dots, v'_n\}$ of $V \subseteq F^{(n)}$, whose respective row matrices are denoted A and A', a *change of base matrix* is a matrix P such that

$$A' = PA. \tag{5-3}$$

Proposition 5.33. *The generalized permutation matrices are the only change of base matrices of s-bases* (*and thus classical bases*).

Proof. Immediate by Corollary 5.25. □

Remark 5.34. It follows from Proposition 5.33, applied to the standard base, that the matrix A is the matrix of a classical base if and only if A is a generalized permutation matrix.

6. Supertropical bilinear forms

The classical way to study orthogonality in vector spaces is by means of bilinear forms. In this section, we introduce the supertropical analog, providing some of the basic properties. Although the tropical literature deals with orthogonality in terms of the inner product, as described in [Akian et al. 2006, §25.6], the supertropical theory leads to a more axiomatic approach.

The notion of supertropical bilinear form follows the classical algebraic theory, although, as is to be expected, there are a few surprises, mostly because of the characteristic 2 nature of the theory [Izhakian et al. 2013]. In this section, we assume that V is a vector space over a supertropical semifield F.

Supertropical bilinear forms.

Definition 6.1. A *(supertropical) bilinear form* on supertropical vector spaces V and V' is a function $B : V \times V' \to F$ satisfying

$$B(v_1 + v_2, w_1 + w_2) \models_{\mathrm{gs}} B(v_1, w_1) + B(v_1, w_2) + B(v_2, w_1) + B(v_2, w_2),$$

$$B(\alpha v, w) = \alpha B(v, w) = B(v, \alpha w),$$

for all $\alpha \in F$, $v_i \in V$, and $w_j \in V'$. We say that a bilinear form B is *strict* if

$$B\big(\alpha_1 v_1 + \alpha_2 v_2, \beta_1 w_1 + \beta_2 w_2\big) \\ = \alpha_1\beta_1 B(v_1, w_1) + \alpha_1\beta_2 B(v_1, w_2) + \alpha_2\beta_1 B(v_2, w_1) + \alpha_2\beta_2 B(v_2, w_2),$$

for all $v_i \in V$ and $w_i \in V'$.

When $V' = V$, we say that B is a *(supertropical) bilinear form* on the vector space V.

We usually write $\langle v, w\rangle$ in place of $B(v, w)$. We do not assume that $\langle v, w\rangle = \langle w, v\rangle$. For the remainder of this section, we take $V' = V \subseteq F^{(n)}$, and consider a (supertropical) bilinear form B on V.

Perhaps surprisingly, one can lift many of the classical theorems about bilinear forms to the supertropical setting, without requiring strictness.

Definition 6.2. The *Gram matrix* of vectors $v_1, \dots, v_k \in V = F^{(n)}$ is defined as the $k \times k$ matrix

$$\widetilde{G}(v_1, \dots, v_k) = \begin{pmatrix} \langle v_1, v_1\rangle & \langle v_1, v_2\rangle & \cdots & \langle v_1, v_k\rangle \\ \langle v_2, v_1\rangle & \langle v_2, v_2\rangle & \cdots & \langle v_2, v_k\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_k, v_1\rangle & \langle v_k, v_2\rangle & \cdots & \langle v_k, v_k\rangle \end{pmatrix}. \tag{6-1}$$

The set $\{v_1, \dots, v_k\}$ is *nonsingular* (with respect to B) if and only if its Gram matrix is nonsingular (see Section 3). The *Gram matrix* of V is the Gram matrix of an s-base of V.

Example 6.3. The quasi-identity

$$\widetilde{G}(v_1, v_2) = \begin{pmatrix} 0 & 1^\nu \\ -\infty & 0 \end{pmatrix} \tag{6-2}$$

(in logarithmic notation) is the Gram matrix of a bilinear form. Note that

$$\langle v_1, v_2\rangle = 1^\nu > 0^\nu = \langle v_1, v_1\rangle + \langle v_2, v_2\rangle.$$

In particular, we have the matrix $\widetilde{G} = \widetilde{G}(b_1 \dots, b_k)$, which can be written as $(g_{i,j})$ where $g_{i,j} = \langle b_i, b_j\rangle$; we describe the bilinear form via the matrix equation

$$\langle v, w\rangle \models_{\mathrm{gs}} v^{\mathrm{t}}\, \widetilde{G} w. \tag{6-3}$$

Of course, the matrix $\widetilde{G}$ depends on the choice of tangible s-base $\mathscr{B}$ of V, but this is unique up to multiplication by scalars and permutation, so $\widetilde{G}$ is unique up to $\widetilde{P}_{\pi;D}\widetilde{G}\widetilde{P}^t_{\pi;D}$, where $\widetilde{P}_{\pi;D}$ is a generalized permutation matrix. In particular, whether or not $\widetilde{G}$ is nonsingular does not depend on the choice of s-base.

Ghost orthogonality. Bilinear forms play a key role in geometry since they permit us to define orthogonality of supertropical vectors. However, as we shall see, orthogonality is rather delicate in this setup. We work with a fixed bilinear form $B = \langle\ ,\ \rangle$ on a supertropical vector space V.

Definition 6.4. For vectors v, w in V, we write $v \perp\!\!\!\perp w$ when $\langle v, w\rangle \in \mathcal{G}_{\mathbb{0}}$, that is, $\langle w_1, w_2\rangle \models_{\text{gs}} \mathbb{0}_F$ (see Remark 2.5), and say that v and w are *left ghost orthogonal*, or *lg-orthogonal* for short. For subspaces W_1, W_2 of V, we say that W_1 is *lg-orthogonal* to W_2 if every $w_1 \in W_1$ is lg-orthogonal to every $w_2 \in W_2$.

The *left orthogonal ghost complement* $S^{\perp\!\!\!\perp}$ of S is defined as

$$S^{\perp\!\!\!\perp} := \{v \in V : \langle v, S\rangle \in \mathcal{G}_{\mathbb{0}}\}.$$

$S^{\perp\!\!\!\perp}$ is always a subspace of V, and $\mathcal{H}_{\mathbb{0}} \subseteq S^{\perp\!\!\!\perp}$ for any $S \subset V$. The *right orthogonal ghost complement* is defined analogously. Of course, when $\langle v, w\rangle \in \mathcal{H}_{\mathbb{0}}$ implies $\langle w, v\rangle \in \mathcal{H}_{\mathbb{0}}$, we may omit the words "left" and "right" and talk of g-orthogonality, which then is a symmetric relation.

Definition 6.5. A subspace W of V is called *nondegenerate* (with respect to B) if $W^{\perp\!\!\!\perp} \cap W \subseteq \mathcal{H}_{\mathbb{0}}$. The bilinear form B is *nondegenerate* if the space V is nondegenerate.

Lemma 6.6. *Suppose $\{w_1, \dots, w_m\}$ tropically spans a subspace W of V, and $v \in V$. If $\sum_{i=1}^{m} \beta_i \langle v, w_i\rangle \in \mathcal{G}_{\mathbb{0}}$ for all $\beta_i \in \mathcal{T}$, then $v \in W^{\perp\!\!\!\perp}$.*

Proof. $\big\langle v, \sum_i \beta_i w_i\big\rangle \models_{\text{gs}} \sum_i \langle v, \beta_i w_i\rangle = \sum_i \beta_i \langle v, w_i\rangle \in \mathcal{G}_{\mathbb{0}}$ for all $\beta_i \in \mathcal{T}$. Thus, $v \in W^{\perp\!\!\!\perp}$. □

Theorem 6.7. *Assume that the vectors $w_1, \dots, w_k \in V$ span a nondegenerate subspace W of V. If $|\widetilde{G}(w_1 \dots, w_k)| \in \mathcal{G}_{\mathbb{0}}$, then $w_1 \dots, w_k$ are tropically dependent.*

Proof. Write $\widetilde{G} = \widetilde{G}(v_1, \dots, v_k)$. By [Izhakian and Rowen 2011a, Theorem 6.6], $|\widetilde{G}| \in \mathcal{G}_{\mathbb{0}}$ if and only if the rows of $\widetilde{G}$ are tropically dependent. By the lemma, if $|\widetilde{G}| \in \mathcal{G}_{\mathbb{0}}$, then some linear combination of the w_i is in $W^{\perp\!\!\!\perp}$. When W is nondegenerate, this latter assertion is the same as saying that the w_i are tropically dependent. □

Corollary 6.8. *If the bilinear form B is nondegenerate on a vector space V, then the Gram matrix* (*with respect to any given supertropical d,s-base of V*) *is nonsingular.*

Remark 6.9. In case the bilinear form B is strict, we can strengthen Lemma 6.6 to obtain, for $v \in V$,

$$v \in W^{\perp\!\!\!\perp} \iff \sum_{i=1}^{m} \beta_i \langle v, w_i\rangle \in \mathcal{G}_{\mathbb{0}} \text{ for all } \beta_i \in \mathcal{T}.$$

(Indeed, if $v \in W^{\perp\!\!\!\perp}$, then $\sum_i \beta_i \langle v, w_i\rangle = \big\langle v, \sum_i \beta_i w_i\big\rangle \in \mathcal{G}_{\mathbb{0}}$ for all i.)

In this case, we can also strengthen Corollary 6.8 to read:

Corollary 6.10. *A strict bilinear form B is nondegenerate on a supertropical vector space V if and only if the Gram matrix* (*with respect to any given supertropical d,s-base of V*) *is nonsingular.*

Symmetry of g-orthogonality.

Definition 6.11. The bilinear form B is *supertropically alternate* if $\langle v, v\rangle \in \mathcal{G}_{\mathbb{0}}$ for all $v \in V$. B is *symmetric* if $\langle v, w\rangle = \langle w, v\rangle$ for all $v, w \in V$. B is *supertropically symmetric* if $\langle v, w\rangle \curlyvee_{\mathrm{gd}} \langle w, v\rangle$ for all $v, w \in V$, that is, $\langle v, w\rangle + \langle w, v\rangle \in \mathcal{G}_{\mathbb{0}}$.

In this subsection, we prove the supertropical version of a classical theorem of Artin, that any bilinear form in which g-orthogonality is symmetric must be a supertropically symmetric bilinear form. (In characteristic 2, any alternate form is symmetric, so we would expect our supertropical forms in the conclusion of the theorem to be symmetric in some sense.)

Definition 6.12. The (supertropical) bilinear form B is *orthogonal-symmetric* if it satisfies the following property for all $v_i, w \in V$:

$$\sum_i \langle v_i, w\rangle \in \mathcal{G}_{\mathbb{0}} \iff \sum_i \langle w, v_i\rangle \in \mathcal{G}_{\mathbb{0}} \tag{6-4}$$

for any finite sum taken over $v_i \in V$.

B is *supertropically orthogonal-symmetric* if B is orthogonal-symmetric and satisfies the additional property that $\langle v, w\rangle \cong_\nu \langle w, v\rangle$ for all $v, w \in V$ satisfying $\langle v, w\rangle \in \mathcal{T}$.

Remark 6.13. If every $\langle v, v\rangle = \mathbb{0}_F$, for all $v \in V$, then the (supertropical) bilinear form B is trivial. (Indeed, $\mathbb{0}_F = \langle v + w, v + w\rangle = \langle v, w\rangle + \langle w, v\rangle$ for all $v, w \in V$, implying $\langle v, w\rangle = \langle w, v\rangle = \mathbb{0}_F$.)

Thus we need to modify our notion of isotropic.

Definition 6.14. A vector $v \in V$ is *g-isotropic* if $\langle v, v\rangle \in \mathcal{G}_{\mathbb{0}}$.

Lemma 6.15. *When the bilinear form B is strict, Condition (6-4) reduces to the condition*

$$\langle v, w\rangle \in \mathcal{G}_{\mathbb{0}} \iff \langle w, v\rangle \in \mathcal{G}_{\mathbb{0}}.$$

Proof. Taking $v = \sum_i v_i$, we have

$$\sum_i \langle v_i, w\rangle = \langle v, w\rangle, \quad \langle w, v\rangle = \sum_i \langle w, v_i\rangle. \qquad \square$$

The symmetry condition extends to sums.

Lemma 6.16. *If B is supertropically symmetric, then*

$$\sum_i \langle v_i, w\rangle \in \mathcal{T} \iff \sum_i \langle w, v_i\rangle \in \mathcal{T}.$$

In this case, $\sum_i \langle v_i, w\rangle = \sum_i \langle w, v_i\rangle$.

Proof. We may assume that $\sum_i \langle v_i, w\rangle \in \mathcal{T}$ and $\sum_i \langle w, v_i\rangle \in \mathcal{T}$, since there is nothing to check if one (and thus the other) is ghost. Take i_1 such that $\langle v_{i_1}, w\rangle$ is the dominant summand of $\sum_i \langle v_i, w\rangle$, and thus is tangible. Likewise, take i_2 such that $\langle w, v_{i_2}\rangle$ is the dominant summand of $\sum_i \langle w, v_i\rangle$, and thus is tangible. By hypothesis, $\langle v_{i_1}, w\rangle = \langle w, v_{i_1}\rangle$ and $\langle w, v_{i_2}\rangle = \langle v_{i_2}, w\rangle$, since these are tangible. Since these dominate their respective sums, we get $\sum_i \langle v_i, w\rangle = \sum_i \langle w, v_i\rangle \in \mathcal{T}$. $\square$

We aim to prove that a g-orthogonal-symmetric (supertropical) bilinear form is supertropically symmetric.

Remark 6.17. The condition $\langle v, w\rangle + \langle w, v\rangle \in \mathcal{G}_{\mathbb{0}}$ means that v is g-orthogonal to w with respect to the new (symmetric) bilinear form given by $\langle v, w\rangle' := \langle v, w\rangle + \langle w, v\rangle$.

Lemma 6.18. *Suppose that B is a g-orthogonal-symmetric bilinear form and $v, w \in V$. Then either $\langle v, w\rangle + \langle w, v\rangle \in \mathcal{G}_{\mathbb{0}}$, or $\langle v, v\rangle = \langle w, w\rangle = \mathbb{0}_F$.*

Proof. One may assume that $\langle v, w\rangle \in \mathcal{T}$; hence $\langle w, v\rangle \in \mathcal{T}$. If $\langle v, w\rangle \cong_\nu \langle w, v\rangle$, then $\langle v, w\rangle + \langle w, v\rangle \in \mathcal{G}_{\mathbb{0}}$, so we may assume by symmetry that $\langle v, w\rangle >_\nu \langle w, v\rangle$.

First assume that w is non-g-isotropic. Then $\gamma\langle v, w\rangle + \langle w, w\rangle$ is ghost for $\gamma = \langle w, w\rangle / \langle v, w\rangle$ and tangible for any other tangible γ in F. At the same time, $\gamma\langle w, v\rangle + \langle w, w\rangle$ is ghost for $\gamma = \langle w, w\rangle / \langle w, v\rangle$, contradicting g-orthogonal-symmetry unless $\langle v, w\rangle \cong_\nu \langle w, v\rangle$. This implies $\langle v, w\rangle + \langle w, v\rangle \in \mathcal{G}_{\mathbb{0}}$.

Next assume that w is g-isotropic but $\langle w, w\rangle = \alpha^\nu \neq \mathbb{0}_F$ for $\alpha \in \mathcal{T}$. Then for tangible $\gamma >_\nu \langle w, w\rangle / \langle v, w\rangle$, we see that $\langle \gamma v, w\rangle + \langle w, w\rangle$ is tangible, so $\langle w, \gamma v\rangle + \langle w, w\rangle$ must also be tangible, which is false if $\gamma <_\nu \langle w, w\rangle / \langle w, v\rangle$. This yields a contradiction if $\langle w, v\rangle <_\nu \langle v, w\rangle$, and similarly we have a contradiction if $\langle w, v\rangle >_\nu \langle v, w\rangle$; hence $\langle w, v\rangle \cong_\nu \langle v, w\rangle$, implying $\langle v, w\rangle + \langle w, v\rangle \in \mathcal{G}_{\mathbb{0}}$.

Thus, we may assume that $\langle w, w\rangle = \mathbb{0}_F$. Likewise, $\langle v, v\rangle = \mathbb{0}_F$, since otherwise we would conclude by interchanging v and w. $\square$

We conclude with our supertropical version of Artin's theorem.

Theorem 6.19. *Every g-orthogonal-symmetric bilinear form B on a supertropical vector space V is supertropically symmetric.*

Proof. We are done by Lemma 6.18 unless there are vectors $v, w \in V$ for which $\langle v, v\rangle = \langle w, w\rangle = \mathbb{0}_F$ and $\langle v, w\rangle + \langle w, v\rangle \in \mathcal{T}$.

In this case, $\alpha := \langle v, w\rangle \in \mathcal{T}$, $\beta := \langle w, v\rangle \in \mathcal{T}$, and $\alpha + \beta \in \mathcal{T}$. Observe that, if $v' \in V$ such that $\langle v', w\rangle \cong_\nu \alpha$, then $\langle w, v'\rangle \cong_\nu \beta$. Indeed, $\langle v, w\rangle + \langle v', w\rangle = \alpha^\nu$, implying $\langle w, v\rangle + \langle w, v'\rangle \in \mathcal{G}_{\mathbb{0}}$. But $\langle w, v'\rangle \in \mathcal{T}$, so we conclude that $\langle w, v'\rangle \cong_\nu \beta$.

Now let vector v' be any vector of V. Then

$$\langle v+v', w\rangle \models_{\mathrm{gs}} \langle v, w\rangle + \langle v', w\rangle \neq \mathbb{0}_F.$$

Thus, $\langle v+v', w\rangle \cong_\nu \gamma$, for some $\gamma \in \mathcal{T}$. Let $v'' := (\alpha/\gamma)(v+v')$. Then

$$\langle v'', w\rangle = \frac{\alpha}{\gamma}\langle v+v', w\rangle \cong_\nu \alpha,$$

and thus $\langle w, v''\rangle \cong_\nu \beta$, as just observed. Hence, $\langle v'', w\rangle + \langle w, v''\rangle \notin \mathcal{G}$. Now Lemma 6.18 yields $\langle v'', v''\rangle = \mathbb{0}_F$. From

$$\mathbb{0}_F = \langle \gamma v'', \gamma v''\rangle \models_{\mathrm{gs}} \langle v, v\rangle + \langle v, v'\rangle + \langle v', v\rangle + \langle v', v'\rangle,$$

we conclude that $\langle v', v'\rangle = \mathbb{0}_F$ for all $v' \in V$; i.e., B is trivial, by Remark 6.13, which is absurd since $\alpha = \langle v, w\rangle \neq \mathbb{0}_F$. Thus, B must be supertropically symmetric. □

Corollary 6.20. *If the bilinear form B is strict and g-orthogonality is a symmetric relation, then B is supertropically symmetric.*

Proof. B is orthogonal-symmetric, by Lemma 6.15. The theorem then can be applied. □

References

[Akian et al. 1990] M. Akian, G. Cohen, S. Gaubert, R. Nikoukhah, and J. P. Quadrat, "Linear systems in (Max-+) algebra", pp. 151–156 in *Proceedings of the 29th Conference on Decision and Control* (Honolulu, 1990), Institute of Electrical and Electronics Engineers, New York, 1990.

[Akian et al. 2006] M. Akian, R. Bapat, and S. Gaubert, "Max-plus algebra", Chapter 25 in *Handbook of linear algebra*, edited by L. Hogben et al., Discrete Mathematics and its Applications **39**, Chapman and Hall/CRC, Boca Raton, FL, 2006.

[Akian et al. 2009] M. Akian, S. Gaubert, and A. Guterman, "Linear independence over tropical semirings and beyond", pp. 1–38 in *Tropical and idempotent mathematics*, edited by G. L. Litvinov and S. N. Sergeev, Contemp. Math. **495**, Amer. Math. Soc., Providence, RI, 2009. MR 2011e:16085 Zbl 1182.15002

[Feichtner and Sturmfels 2005] E. M. Feichtner and B. Sturmfels, "Matroid polytopes, nested sets and Bergman fans", *Port. Math.* (*N.S.*) **62**:4 (2005), 437–468. MR 2006j:05036 Zbl 1092.52006 arXiv math/0411260v1

[Gaubert and Katz 2007] S. Gaubert and R. D. Katz, "The Minkowski theorem for max-plus convex sets", *Linear Algebra Appl.* **421**:2-3 (2007), 356–369. MR 2007k:52001 Zbl 1110.52002

[Izhakian 2009] Z. Izhakian, "Tropical arithmetic and matrix algebra", *Comm. Algebra* **37**:4 (2009), 1445–1468. MR 2010d:16059 Zbl 1165.15017 arXiv math.AG/0505458

[Izhakian and Rowen 2009] Z. Izhakian and L. Rowen, "The tropical rank of a tropical matrix", *Comm. Algebra* **37**:11 (2009), 3912–3927. MR 2010k:15003 Zbl 1184.15003

[Izhakian and Rowen 2010] Z. Izhakian and L. Rowen, "Supertropical algebra", *Adv. Math.* **225**:4 (2010), 2222–2286. MR 2012a:14137 Zbl 05796760 arXiv 0806.1171

[Izhakian and Rowen 2011a] Z. Izhakian and L. Rowen, "Supertropical matrix algebra", *Israel J. Math.* **182** (2011), 383–424. MR 2012e:16108 Zbl 1215.15018

[Izhakian and Rowen 2011b] Z. Izhakian and L. Rowen, "Supertropical matrix algebra II: Solving tropical equations", *Israel J. Math.* **186** (2011), 69–96. MR 2852317 Zbl 06108394 arXiv 0902.2159

[Izhakian and Rowen 2011c] Z. Izhakian and L. Rowen, "Supertropical matrix algebra, III: Powers of matrices and their supertropical eigenvalues", *J. Algebra* **341** (2011), 125–149. MR 2012e:15054 Zbl 06005428 arXiv 1008.0023

[Izhakian et al. 2011] Z. Izhakian, M. Johnson, and M. Kambites, "Pure dimension and projectivity of tropical convex sets", preprint, 2011. arXiv 1106.4525

[Izhakian et al. 2013] Z. Izhakian, M. Knebusch, and L. Rowen, "Supertropical quadratic forms, I", preprint, 2013. arXiv 1309.5729

[Maclagan and Sturmfels 2009] D. Maclagan and B. Sturmfels, "Introduction to tropical geometry", 2009, http://homepages.warwick.ac.uk/staff/D.Maclagan/papers/TropicalBook.html.

Received August 4, 2010. Revised November 20, 2012.

ZUR IZHAKIAN
DEPARTMENT OF MATHEMATICS
BAR-ILAN UNIVERSITY
RAMAT-GAN 52900
ISRAEL
zzur@math.biu.ac.il

MANFRED KNEBUSCH
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF REGENSBURG
REGENSBURG
GERMANY
manfred.knebusch@mathematik.uni-regensburg.de

LOUIS ROWEN
DEPARTMENT OF MATHEMATICS
BAR-ILAN UNIVERSITY
RAMAT-GAN 52900
ISRAEL
rowen@macs.biu.ac.il

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.77

ISOMETRY GROUPS AMONG TOPOLOGICAL GROUPS

PIOTR NIEMIEC

It is shown that a topological group G is topologically isomorphic to the isometry group of a (complete) metric space if and only if G coincides with its $\mathscr{G}_\delta$-closure in the Raĭkov completion of G (resp. if G is Raĭkov-complete). It is also shown that for every Polish (resp. compact Polish; locally compact Polish) group G there is a complete (resp. proper) metric d on X inducing the topology of X such that G is isomorphic to $\mathrm{Iso}(X, d)$, where $X = \ell_2$ (resp. $X = [0, 1]^\omega$; $X = [0, 1]^\omega \setminus \{\text{point}\}$). It is demonstrated that there are a separable Banach space E and a nonzero vector $e \in E$ such that G is isomorphic to the group of all (linear) isometries of E which leave the point e fixed. Similar results are proved for arbitrary Raĭkov-complete topological groups.

1. Introduction

Gao and Kechris [2003] proved that every Polish group is isomorphic to the (full) isometry group of some separable complete metric space. Melleray [2008] and Malicki and Solecki [2009] improved this result in the context of compact and, respectively, locally compact Polish groups by showing that every such group is isomorphic to the isometry group of a compact and, respectively, a proper metric space. (A metric space is *proper* if and only if each closed ball in this space is compact). All their proofs were complicated and based on the techniques of the so-called *Katětov* maps. In [Niemiec 2012] we introduced a new method to characterize groups of homeomorphisms of a locally compact Polish space which coincide with the isometry groups of the space with respect to some proper metrics. As a consequence, we showed that every (separable) Lie group is isomorphic to the isometry group of another Lie group equipped with some proper metric and that every finite-dimensional [locally] compact Polish group is isomorphic to the isometry group of a finite-dimensional [proper locally] compact metric space. One

MSC2010: primary 22A05, 54H11; secondary 57N20.
Keywords: Polish group, isometry group, Hilbert cube, Hilbert space, Hilbert cube manifold, Raĭkov-complete group, isometry group of a Banach space.

of the aims of this paper is to give the results of Gao and Kechris, Melleray, and Malicki and Solecki a more "explicit" and unified form:

Theorem 1.1. *Let G be a Polish group.*

(a) *There is a complete compatible metric d on ℓ_2 such that G is isomorphic to* $\mathrm{Iso}(\ell_2, d)$.

(b) *If G is compact, there is a compatible metric d on the Hilbert cube Q such that G is isomorphic to* $\mathrm{Iso}(Q, d)$.

(c) *If G is locally compact, there is a proper compatible metric d on* $Q \setminus \{\text{point}\}$ *such that G is isomorphic to* $\mathrm{Iso}(Q \setminus \{\text{point}\}, d)$.

We shall also prove the following:

Theorem 1.2. *For every Polish group G there exist a separable real Banach space E and a nonzero vector $e \in E$ such that G is isomorphic to the group of all linear isometries of E (endowed with the pointwise convergence topology) which leave the point e fixed.*

Our methods can be adapted to general settings and give a characterization of topological groups which are isomorphic to isometry groups of complete as well as incomplete metric spaces. To this end, we recall that a topological group G is *Raĭkov-complete* (or *upper-complete*) if and only if it is complete with respect to the upper uniformity, by [Arhangel'skii and Tkachenko 2008, §3.6] or [Roelcke and Dierolf 1981] (see also the remarks on page 1581 in [Uspenskij 2008]). In other words, G is upper-complete if every net $\{x_\sigma\}_{\sigma \in \Sigma} \subset G$ satisfying the following condition is convergent in G:

(C) For every neighborhood U of the neutral element of G there is $\sigma_0 \in \Sigma$ such that both $x_\sigma x_{\sigma'}^{-1}$ and $x_\sigma^{-1} x_{\sigma'}$ belong to U for any $\sigma, \sigma' \geqslant \sigma_0$.

Equivalently, the net $\{x_\sigma\}_{\sigma \in \Sigma}$ satisfies (C) if both the nets $\{x_\sigma\}_{\sigma \in \Sigma}$ and $\{x_\sigma^{-1}\}_{\sigma \in \Sigma}$ are fundamental with respect to the left uniformity of G. We call Raĭkov-complete groups briefly *complete*, following [Uspenskij 2008]. The class of all complete topological groups coincides with the class of all absolutely closed topological groups (a topological group is *absolutely closed* if it is closed in every topological group containing it as a topological subgroup). It is well-known that for every topological group G there exists a unique (up to topological isomorphism) complete topological group containing G as a dense subgroup (see, e.g., [Roelcke and Dierolf 1981; Arhangel'skii and Tkachenko 2008, §3.6]). This complete group is called the *Raĭkov completion* of G and we shall denote it by $\overline{G}$.

Less classical are topological groups, which we call $\mathscr{G}_\delta$-complete. To define them, let us agree with the following general convention: Whenever τ is a topology on a set X, τ_δ stands for the topology on X whose base is formed by all $\mathscr{G}_\delta$-sets (with respect to τ) in X. (In particular, $\mathcal{G}_\delta$-sets in (X, τ) are open in (X, τ_δ).) Subsets of X which are closed or dense in the topology τ_δ are called $\mathcal{G}_\delta$-closed and $\mathcal{G}_\delta$-dense, respectively (see, for example, [Arhangel'skii 2002; Arhangel'skii and Tkachenko 2008, page 268]).

It may be easily verified that if (G, τ) is a topological group, so is (G, τ_δ).

Definition 1.3. A topological group G is $\mathcal{G}_\delta$*-complete* if (G, τ_δ) is a complete topological group (where τ is the topology of G).

Equivalently, a topological group G is $\mathcal{G}_\delta$-complete if and only if G is $\mathcal{G}_\delta$-closed in $\overline{G}$. The class of all $\mathcal{G}_\delta$-complete groups is huge (see Proposition 4.3 below) and contains all complete as well as metrizable topological groups (more detailed discussion on this class is included in Section 4). (It is worth noting here that, according to the Birkhoff–Kakutani theorem, a topological group is metrizable if and only if it is first-countable, that is, if it has a countable base of neighborhoods of the neutral element. For a proof see, for example, Theorem 3.3.12 in [Arhangel'skii and Tkachenko 2008].) However, there are topological groups which are not $\mathcal{G}_\delta$-complete (see Example 4.5 below).

$\mathcal{G}_\delta$-complete groups turn out to characterize isometry groups of metric spaces:

Theorem 1.4. *Let G be a topological group*:

(A) *The following conditions are equivalent*:

  (A1) *There exists a metric space* (X, d) *such that G is isomorphic to* $\mathrm{Iso}(X, d)$.

  (A2) *G is $\mathcal{G}_\delta$-complete.*

  Moreover, *if G is $\mathcal{G}_\delta$-complete*, *the space X witnessing* (A1) *may be chosen so that $w(X) = w(G)$.*

(B) *The following conditions are equivalent*:

  (B1) *There exists a complete metric space* (X, d) *such that G is isomorphic to* $\mathrm{Iso}(X, d)$.

  (B2) *G is complete.*

  Moreover, *if G is complete*, *the space X witnessing* (B1) *may be chosen so that $w(X) = w(G)$.*

(By $w(X)$ we denote the topological weight of a topological space X.)

One concludes from Theorem 1.4 that the isometry group of an arbitrary metric space is always Dieudonné-complete (see Corollary 4.4 below). This solves a problem posed, for example, by Arhangel'skii and Tkachenko [2008, Open Problem 3.5.4 on page 181].

A generalization of Theorems 1.1 and 1.2 has the following form:

Proposition 1.5. *Let G be a complete topological group of topological weight not greater than $\beta \geqslant \aleph_0$.*

(a) *There is a complete compatible metric ϱ on $\mathcal{H}_\beta$ such that G is isomorphic to* $\mathrm{Iso}(\mathcal{H}_\beta, \varrho)$, *where $\mathcal{H}_\beta$ is a real Hilbert space of* (*Hilbert space*) *dimension equal to β.*

(b) *There are an infinite-dimensional real Banach space E of topological weight β and a nonzero vector $e \in E$ such that G is isomorphic to the group of all linear isometries of E which leave the point e fixed.*

As an immediate consequence of Theorem 1.4 and Proposition 1.5 we obtain:

Corollary 1.6. *Let $\mathcal{H}$ be a Hilbert space of Hilbert space dimension $\beta \geqslant \aleph_0$ and let*

$$\mathcal{G} = \{\mathrm{Iso}(\mathcal{H}, \varrho) \mid \varrho \text{ is a complete compatible metric on } \mathcal{H}\}.$$

Then, *up to isomorphism*, *$\mathcal{G}$ consists precisely of all complete topological groups of topological weight not exceeding β.*

The paper is organized as follows: In Section 2 we give a new proof of the Gao–Kechris theorem mentioned above. We consider our proof more transparent, more elementary, and less complicated. The techniques of this part are adapted in Section 3, where we demonstrate that every closed subgroup of the isometry group of a (complete) metric space (X, d) is actually (isomorphic to) the isometry group of a certain (complete) metric space, closely “related” to (X, d). This theorem is applied in Section 4, where we establish basic properties of the class of all $\mathcal{G}_\delta$-complete groups and prove Theorem 1.4. Section 5 contains proofs of Theorem 1.2, Proposition 1.5, Theorem 1.1(a), and Corollary 1.6. In Section 6 we study topological groups isomorphic to isometry groups of completely metrizable metric spaces. Section 7 is devoted to the proofs of points (b) and (c) of Theorem 1.1.

Notation and terminology. In this paper $\mathbb{N} = \{0, 1, 2, \dots\}$ (and it is equipped with the discrete topology). All isomorphisms between topological groups are topological, all topological groups are Hausdorff, and all isometries between metric spaces are, by definition, bijective. All normed vector spaces are assumed to be real. The topological weight of a topological space X is denoted by $w(X)$ and it is understood as an infinite cardinal number. Isometry groups (and all their subsets) of metric as well as normed vector spaces are endowed with the pointwise convergence topology, which makes them topological groups. A *Polish* space (resp. group) is a completely metrizable separable topological space (resp. group). A metric on a topological space is *compatible* if and only if it induces the topology of the space. It is *proper* if all closed balls with respect to this metric are compact (in the topology induced by this metric). Whenever (X, d) is a metric space, $a \in X$ and $r > 0$, $B_X(a, r)$ and $\bar{B}_X(a, r)$ stand for, respectively, the open and the closed d-balls with center at a and of radius r. The Hilbert cube, that is, the countable infinite Cartesian power of [0, 1], is denoted by Q and ℓ_2 stands for the separable Hilbert space. A *map* means a continuous function.

2. The Gao–Kechris theorem revisited

This part is devoted to the proof of the Gao–Kechris theorem [2003] mentioned in the introductory part and stated below. Another proof may be found in [Melleray 2008].

Theorem 2.1. *Every Polish group is isomorphic to the isometry group of a certain separable complete metric space.*

For the purpose of this and the next section, let us agree with the following conventions: For every nonempty collection $\{X_s\}_{s\in S}$ of topological spaces, $\bigsqcup_{s\in S} X_s$ denotes the topological disjoint union of these spaces. In particular, whenever the notation $\bigsqcup_{s\in S} X_s$ appears, the sets $X_s (s \in S)$ are assumed to be pairwise disjoint (the same rule for the symbol "$\sqcup$"). For a function $f : X \to X$ and an integer $n \geqslant 1$, we denote by $f^{ⓝ} : X^n \to X^n$ the n-th Cartesian power of f, given by $f^{ⓝ}(x_1, \dots, x_n) = (f(x_1), \dots, f(x_n))$; and by $f \times w$, for an arbitrary point w, we denote the map $X \times \{w\} \to X \times \{w\}$ that sends (x, w) to $(f(x), w)$ for any $x \in X$. Similarly, if d is a metric on X, we denote by $d^{ⓝ}$ the maximum metric on X^n induced by d; that is,

$$d^{ⓝ}((x_1, \dots, x_n), (y_1, \dots, y_n)) = \max_{j=1,\dots,n} d(x_j, y_j),$$

and $d \times w$ is the metric on $X \times \{w\}$ such that $(d \times w)((x, w), (y, w)) = d(x, y)$. Finally, for a topological space V and a map $v : V \to V$ we put $\widehat{V} = (V \times \mathbb{N}) \sqcup \left(\bigsqcup_{n=2}^{\infty} V^n\right) \sqcup \mathbb{N}$ and define $\hat{v} : \widehat{V} \to \widehat{V}$ by the following rules: $\hat{v}(x, m) = (v(x), m)$, $\hat{v}|_{V^n} = v^{ⓝ}$, and $\hat{v}(m) = m$ for any $x \in V$, $m \in \mathbb{N}$, and $n \in \mathbb{N} \setminus \{0\}$. To avoid repetitions, for a metric space (X, d) and arbitrary sets $A, B \subset \mathbb{N}$, and $C \subset \mathbb{N} \setminus \{0, 1\}$, let us say a metric ϱ on $(X \times A) \sqcup \left(\bigsqcup_{j\in C} X^j\right) \sqcup B$ $(\subset \widehat{X})$ *respects* d if and only if the following three conditions are satisfied:

(AX1) ϱ coincides with $d \times m$ on $X \times \{m\}$ for each $m \in A$.

(AX2) ϱ coincides with $d^{ⓝ}$ on X^n for each $n \in C$.

(AX3) $\varrho(x, y) \geqslant 1$ whenever x and y belong to distinct members of the collection $\{X \times \{m\} \mid m \in A\} \cup \{X^n \mid n \in C\} \cup \{\{k\} \mid k \in B\}$.

Observe that (AX1)–(AX3) imply that

(AX4) if ϱ respects d then ϱ is compatible; if, moreover, d is complete, so is ϱ.

The main result of this section is the following:

Proposition 2.2. *Let (X, d) be a separable bounded complete metric space and let G be a closed subgroup of* $\mathrm{Iso}(X, d)$. *There exists a metric ϱ on $\widehat{X}$ such that ϱ respects d and the function*

$$G \ni u \mapsto \hat{u} \in \mathrm{Iso}(\widehat{X}, \varrho)$$

is a well-defined isomorphism of topological groups.

The proof of Proposition 2.2 will be preceded by a few auxiliary results. The first of them is a kind of folklore and we leave its (simple) proof to the reader.

Lemma 2.3. *Let $\{(X_s, d_s)\}_{s\in S}$ be a nonempty family of metric spaces such that for $A = \bigcap_{s\in S} X_s$ we have:*

- *$X_s \cap X_{s'} = A$ and $d_s|_{A\times A} = d_{s'}|_{A\times A}$ for any two distinct indices s and s' of S.*
- *A is nonempty and closed in (X_s, d_s) for each $s \in S$.*

Let $X = \bigcup_{s\in S} X_s$ and let $d : X \times X \to [0, \infty)$ be given by the rules:

- *d coincides with d_s on $X_s \times X_s$ for every $s \in S$.*
- *$d(x, y) = \inf\{d_s(x, a) + d_{s'}(a, y) \mid a \in A\}$ whenever $x \in X_s \setminus X_{s'}$ and $y \in X_{s'} \setminus X_s$ for distinct indices s and s'.*

Then d is a well-defined metric on X with the following property: Whenever $f_s \in$ $\mathrm{Iso}(X_s, d_s)(s \in S)$ are maps such that $f_s|_A = f_{s'}|_A$ and $f_s(A) = A$ for any $s, s' \in S$ then their union $f := \bigcup_{s\in S} f_s$ (that is, $f = f_s$ on X_s) is a well-defined function such that $f \in \mathrm{Iso}(X, d)$.

The above result will be the main tool for constructing the metric ϱ appearing in Proposition 2.2.

In the next two results, (X, p) is a complete nonempty metric space with

$$p < 1. \tag{1}$$

Lemma 2.4. *Let $J \subset \mathbb{N} \setminus \{0\}$ be a finite set such that $n = \mathrm{card}(J) > 1$. There is a metric λ on $F := [X \times (J \cup \{0\})] \sqcup X^n$ with the following properties:*

(a) *λ respects p and $\lambda \leqslant 5$.*

(b) *For every $u \in \mathrm{Iso}(X, p)$, $\hat{u}|_F \in \mathrm{Iso}(F, \lambda)$.*

(c) *If $g \in \mathrm{Iso}(F, \lambda)$ is such that $g(X \times \{j\}) = X \times \{j\}$ for each $j \in J \cup \{0\}$ then $g = \hat{u}|_F$ for some $u \in \mathrm{Iso}(X, p)$.*

Proof. With no loss of generality, we may assume that $J = \{1, \dots, n\}$. Let $A = \{(x_1, \dots, x_n) \in X^n \mid x_1 = \cdots = x_n\}$ and let λ_0' be the metric on $X_0{}' :=$ $(X \times \{0\}) \sqcup A$ that coincides with $p \times 0$ on $X \times \{0\}$, with $p^{\textcircled{n}}$ on A, and such that $\lambda_0'((x, 0), (a, \dots, a)) = 1 + p(x, a)$ for $x \in X$ and $(a, \dots, a) \in A$. Now apply Lemma 2.3 for $\{(X_0{}', \lambda_0'), (X^n, p^{\textcircled{n}})\}$ to obtain a metric λ_0 on $X_0 := (X \times \{0\}) \sqcup X^n$

which extends both λ_0' and $p^{ⓝ}$. Observe that λ_0 respects p, that $\lambda_0 \leqslant 3$ (by (1)), that $\hat{u}|_{X_0} \in \mathrm{Iso}(X_0, \lambda_0)$ for each $u \in \mathrm{Iso}(X, p)$, and that, for arbitrary $x, x_1, \dots, x_n \in X$,

$$\lambda_0((x, 0), (x_1, \dots, x_n)) = 1 \iff x_1 = \cdots = x_n = x. \tag{2}$$

Further, for $j \in J$ let λ_j be the metric on $X_j := (X \times \{j\}) \sqcup X^n$ that coincides with $p \times j$ on $X \times \{j\}$, with $p^{ⓝ}$ on X^n, and such that $\lambda_j((x, j), (x_1, \dots, x_n)) = 1 + p(x, x_j)$ for any $x, x_1, \dots, x_n \in X$ (λ_j is indeed a metric thanks to (1)). Similarly, as before, notice that λ_j respects p, $\lambda_j \leqslant 2$ and for any $x, x_1, \dots, x_n \in X$:

$$\lambda_j((x, j), (x_1, \dots, x_n)) = 1 \iff x_j = x. \tag{3}$$

Now again apply Lemma 2.3 for the family $\{(X_j, \lambda_j) \mid j \in J \cup \{0\}\}$ to obtain a metric λ on F which extends each of $\lambda_j (j \in J \cup \{0\})$. It follows from the construction and Lemma 2.3 that points (a) and (b) are satisfied. We turn to (c). Let g be as specified there. Let $u : X \to X$ be such that $u \times 0 = g|_{X \times \{0\}}$ and, similarly, for $j \in J$ let $u_j : X \to X$ be such that $u_j \times j = g|_{X \times \{j\}}$. Finally, put $f = g|_{X^n} : X^n \to X^n$. Since λ respects p, $u \in \mathrm{Iso}(X, p)$. So we only need to check that $u_1 = \dots u_n = u$ and $f = u^{ⓝ}$. Let $\pi_j : X^n \to X$ be the projection onto the j-th coordinate ($j = 1, \dots, n$). For any $x = (x_1, \dots, x_n) \in X^n$ and $j \in J$ we have, by (3),

$$1 = \lambda((\pi_j(x), j), x) = \lambda(g(\pi_j(x), j), g(x)) = \lambda(((u_j \circ \pi_j)(x), j), f(x))$$

and therefore, again by (3), $u_j \circ \pi_j = \pi_j \circ f$. Consequently, $f(x_1, \dots, x_n) = (u_1(x_1), \dots, u_n(x_n))$. Finally, for any $z \in X$ we have, by (2),

$$1 = \lambda((z, 0), (z, \dots, z)) = \lambda(g(z, 0), g(z, \dots, z)) = \lambda((u(z), 0), (u_1(z), \dots, u_n(z)))$$

and hence, again by (2), $u_1(z) = \cdots = u_n(z) = u(z)$. □

Lemma 2.5. *Let G be a subgroup of* $\mathrm{Iso}(X, p)$ *and let $z \in X^n$ and $J \subset \mathbb{N} \setminus \{0\}$ be such that* $\mathrm{card}(J) = n > 1$. *Let D denote the closure* (*in X^n*) *of the set* $\{u^{ⓝ}(z) \mid u \in G\}$. *There exists a metric μ on $F := [X \times (J \cup \{0\})] \sqcup X^n \sqcup \{n-1\}$ which has the following properties*:

(a) *μ respects p and $\mu \leqslant 11$.*

(b) *$\hat{u}|_F \in \mathrm{Iso}(F, \mu)$ for every $u \in G$.*

(c) *For any $g \in \mathrm{Iso}(F, \mu)$ there is $u \in \mathrm{Iso}(X, p)$ such that $g = \hat{u}|_F$ and $u^{ⓝ}(z) \in D$.*

Proof. Without loss of generality, we may assume that $J = \{1, \dots, n\}$. Let λ be as in Lemma 2.4 (so, λ is a metric on $F \setminus \{n-1\}$). Let $c_0, \dots, c_{n+1}$ be such that

$$5 < c_0 < c_1 < \cdots < c_{n+1} < 6. \tag{4}$$

Put $A = [X \times (J \cup \{0\})] \sqcup D$ and denote by μ_0 the metric on $A \sqcup \{n-1\}$ such that μ_0 coincides with λ on A, $\mu_0((x, j), n-1) = c_j$ for $x \in X$ and $j = 0, \dots, n$, and $\mu_0(y, n-1) = c_{n+1}$ for $y \in D$ (μ_0 is a metric thanks to Lemma 2.4(a), (AX3), and (4)). Now apply Lemma 2.3 for the family $\{(A \cup \{n-1\}, \mu_0), (F \setminus \{n-1\}, \lambda)\}$ to obtain a metric μ which extends both μ_0 and λ. We infer the validity of (a) from (4) and Lemma 2.4(a). Further, since $u^{ⓝ}(D) = D$ for each $u \in G$ and thanks to of Lemma 2.4(b), condition (b) is fulfilled as well (see Lemma 2.3). We turn to (c).

Let $g \in \mathrm{Iso}(F, \mu)$. Since $n-1$ is a unique point $q \in F$ such that $\mu(q, x) = c_0$ and $\mu(q, y) = c_1$ for some $x, y \in F$ (since $\lambda \leqslant 5 < c_0 < c_1$), we conclude that $g(n-1) = n-1$. Further, observe that for each $x \in X^n$, $\mu(x, n-1) \geqslant c_{n+1}$ because of (AX3) and (4). Consequently, $X \times \{j\} = \{x \in F \mid \mu(x, n-1) = c_j\}$ for $j = 0, 1, \dots, n$. Thus, we see that $g(X \times \{j\}) = X \times \{j\}$ for such j's. Since $g|_{F\setminus\{n-1\}} \in \mathrm{Iso}(F \setminus \{n-1\}, \lambda)$ and $g(n-1) = n-1$, point (c) of Lemma 2.4 implies that there is $u \in \mathrm{Iso}(X, p)$ such that $g = \hat{u}|_F$. Finally, $g(z) = u^{ⓝ}(z) \in X^n$, and for $y \in X^n$, $\mu(y, n-1) = c_{n+1}$ if and only if $y \in D$ (by (4) and (AX3)), which gives $u^{ⓝ}(z) \in D$. □

Proof of Proposition 2.2. Let $r \geqslant 1$ be such that $d < r$. Put $p = d/r < 1$ and notice that $\mathrm{Iso}(X, p) = \mathrm{Iso}(X, d)$. Let $X_0 = \{x_n \mid n \geqslant 1\}$ be a dense subset of X. Let $J_1, J_2, \dots$ be pairwise disjoint sets such that $\bigcup_{n=1}^{\infty} J_n = \mathbb{N} \setminus \{0\}$ and $\mathrm{card}(J_n) = n+1$. For each $n \geqslant 2$ put $z_n = (x_1, \dots, x_n) \in X^n$, $F_n = [X \times (J_{n-1} \cup \{0\})] \sqcup X^n \sqcup \{n-1\}$, and let D_n be the closure (in X^n) of $\{u^{ⓝ}(z_n) \mid u \in G\}$. Further, let μ_n be a metric on F_n obtained from Lemma 2.5 (applied for z_n and J_{n-1}). Now apply Lemma 2.3 for the collection $\{(F_n, \mu_n) \mid n \geqslant 2\}$ to get a metric λ_0 on $\widehat{X} \setminus \{0\}$ which extends each of μ_n ($n \geqslant 2$). In particular, λ_0 respects p and $\lambda_0 \leqslant 22$. Finally, we extend the metric λ_0 to a metric λ on $\widehat{X}$ in such a way that for $k \geqslant 0$, $\lambda(x, 0) = c_{k,1}$ for $x \in X \times \{k\}$, $\lambda(x, 0) = c_{k,2}$ for $x \in X^{k+2}$, and $\lambda(k+1, 0) = c_{k,3}$, where

$$c_{0,1}, c_{0,2}, c_{0,3}, c_{1,1}, c_{1,2}, c_{1,3}, \dots \text{ are all different numbers,} \tag{5}$$

are greater than 22, and smaller than 23 (λ is a metric thanks to (AX3)). It follows from Lemma 2.3 and Lemma 2.5(b) that $\hat{u} \in \mathrm{Iso}(\widehat{X}, \lambda)$ for any $u \in G$. It is clear that the function $G \ni u \mapsto \hat{u} \in \mathrm{Iso}(\widehat{X}, \lambda)$ is a group homomorphism and a topological embedding. We shall now show that it is also surjective.

Let $g \in \mathrm{Iso}(\widehat{X}, \lambda)$. Since 0 is a unique point $q \in \widehat{X}$ such that $\lambda(q, x) = c_{0,1}$ and $\lambda(q, y) = c_{0,2}$ for some $x, y \in \widehat{X}$, we see that $g(0) = 0$. Consequently, $g(X \times \{k\}) = X \times \{k\}$, $g(X^{k+2}) = X^{k+2}$, and $g(k+1) = k+1$ for each $k \geqslant 0$, by (5). So, taking into account that $g|_{F_n} \in \mathrm{Iso}(F_n, \mu)$, point (c) of Lemma 2.5 yields that there is $u \in \mathrm{Iso}(X, p)$ such that $g = \hat{u}$ and $u(z_n) \in D_n$. The latter condition implies that there are elements $u_1, u_2, \dots$ of G which converge pointwise to u on X_0. We now infer from the density of X_0 in X that $u = \lim_{n\to\infty} u_n$, and in fact $u \in G$ by the closedness of G.

To end the proof, it suffices to put $\varrho = r\lambda$. □

Proof of Theorem 2.1. Let $(H, \cdot)$ be a Polish group. First we introduce a standard argument used, for example, by Melleray [2008] in his proof of this theorem: Take a left-invariant metric $d_0 \leqslant 1$ on H and denote by (X, d) the completion of (H, d_0). Then, of course, X is separable and for every $h \in H$ there is a unique $u_h \in \mathrm{Iso}(X, d)$ such that $u_h(x) = hx$ for $x \in H$. Observe that the function $H \ni h \mapsto u_h \in \mathrm{Iso}(X, d)$ is a group homomorphism as well as a topological embedding. Therefore, its image G is isomorphic to H. Since G is a Polish subgroup of a Polish group, G is closed in $\mathrm{Iso}(X, d)$. Now it suffices to apply Proposition 2.2 and to use (AX4) to deduce the completeness of the metric obtained by that result. □

3. Closed subgroups of isometry groups

In this section we generalize the ideas of the previous part to the context of all isometry groups. Our aim is to show that a closed subgroup of the isometry group of a metric space is isomorphic to the isometry group of another metric space. We have decided to discuss the separable case separately, because in that case the proofs are more transparent and easier. Actually all tools were prepared in the previous section, except the following one:

Lemma 3.1. *Let X be a set with* $\mathrm{card}(X) \neq 2$ *and $I \subset (0, \infty)$ be a nondegenerate interval. There is a metric $d : X \times X \to I \cup \{0\}$ such that the identity map of X is a unique member of* $\mathrm{Iso}(X, d)$.

We shall prove a stronger version of Lemma 3.1 at the end of the section. Now we generalize the concepts in Section 3. Let β be an infinite cardinal number and let D_β denote a fixed discrete topological space of cardinality β. For a metrizable space X and a function $f : X \to X$ let X^0 be a one-point space and $f^{⓪} : X^0 \to X^0$ denote the identity map. Further, we put $T(X) = \bigsqcup_{n \in \mathbb{N}} X^n$ (recall that $0 \in \mathbb{N}$). Finally, denote by $\widehat{X}_\beta$ and $\hat{f}_\beta$ (resp.) the product $T(X) \times D_\beta$ and the function of $\widehat{X}_\beta$ into itself such that $\hat{f}_\beta = f^{ⓝ} \times \xi$ on $X^n \times \{\xi\}$ for any $n \in \mathbb{N}$ and $\xi \in D_\beta$. (Notice that $w(\widehat{X}_\beta) = \beta$ provided $\beta \geqslant w(X)$.) For any $J \subset \mathbb{N}$ and a collection $\{A_n\}_{n \in J}$ of subsets of D_β, we say a metric ϱ on $\bigsqcup_{n \in J}(X^n \times A_n) \subset \widehat{X}_\beta$ *respects* a compatible metric d on X if and only if the following two conditions are fulfilled:

(PR1) ϱ coincides with $d^{ⓝ} \times \xi$ on $X^n \times \{\xi\}$ for any $n \in J \setminus \{0\}$ and $\xi \in A_n$.

(PR2) $\varrho(x, y) \geqslant 1$ whenever x and y belong to distinct members of the collection

$$\{X^n \times \{\xi\} \mid n \in J,\ \xi \in A_n\}.$$

As before, we see that (AX4) is satisfied.

A counterpart of Proposition 2.2 in the general case is the following:

Theorem 3.2. *Let β be an infinite cardinal number and (X,d) be a nonempty bounded metric space such that $w(X) \leqslant \beta$. For any closed subgroup G of* $\mathrm{Iso}(X,d)$ *there exists a metric ϱ on $\widehat{X}_\beta$ such that ϱ respects d and the function*

$$G \ni u \mapsto \hat{u}_\beta \in \mathrm{Iso}(\widehat{X}_\beta, \varrho) \tag{6}$$

is a well-defined isomorphism of topological groups.

Proof. In what follows, we shall (naturally) identify $X^0 \times D_\beta$ with D_β. It follows from the proof of Proposition 2.2 that we may assume $d < 1$. Let Z be a dense set in X such that $\mathrm{card}(Z) \leqslant \beta$. Fix arbitrary $\theta \in D_\beta$ and write $D_\beta \setminus \{\theta\}$ in the form $\bigcup_{n=0}^{\infty} S_n$, where $\mathrm{card}(S_n) = \beta$ for any n and

$$S_n \cap S_m = \varnothing \quad (n \neq m). \tag{7}$$

The set S_0 and the point θ will be employed in the last part of the proof. For simplicity, put $S_* = \bigcup_{n=1}^{\infty} S_n$ and $X_* := \widehat{X}_\beta \setminus (S_0 \cup \{\theta\}) = \big(\bigsqcup_{n\geqslant 1}(X^n \times D_\beta)\big) \sqcup S_*$. It follows from (7) that for any $\xi \in D_\beta \setminus \{\theta\}$ there is a unique number $n(\xi) \in \mathbb{N}$ such that $\xi \in S_{n(\xi)-1}$. Further, for every $n \geqslant 2$ there are a surjection $\kappa_n : S_{n-1} \to Z^n$ and a bijection $\tau_n : S_{n-1} \to D_\beta$. Take a collection $\{J_\xi \mid \xi \in S_*\}$ such that for any $\xi, \eta \in S_*$

(S1) $\mathrm{card}(J_\xi) = n(\xi) (\geqslant 2)$;

(S2) $J_\xi \cap J_\eta = \varnothing$ whenever $\xi \neq \eta$;

(S3) $\bigcup_{\zeta \in S_*} J_\zeta = D_\beta \setminus \{\theta\}$.

We deduce from (S1) and Lemma 2.5 that for each $\xi \in S_*$ there exists a metric μ_ξ on $F_\xi := [X \times (J_\xi \cup \{\theta\})] \sqcup (X^{n(\xi)} \times \{\tau_{n(\xi)}(\xi)\}) \sqcup \{\xi\}$ which has the following properties:

(D1) μ_ξ respects d and $\mu_\xi \leqslant 11$.

(D2) $\hat{u}_\beta|_{F_\xi} \in \mathrm{Iso}(F_\xi, \mu_\xi)$ for every $u \in G$.

(D3) For any $g \in \mathrm{Iso}(F_\xi, \mu_\xi)$ there is $u \in \mathrm{Iso}(X,d)$ with $g = \hat{u}_\beta|_{F_\xi}$, and $u^{ⓝ}(\kappa_n(\xi))$ belongs to the closure B_ξ of $\{f^{ⓝ}(\kappa_n(\xi)) \mid f \in G\}$ in X^n, where $n = n(\xi)$.

Observe that (S2)–(S3), (D1), and the bijectivity of the τ_n imply that

$$F_\xi \cap F_\eta = X \times \{\theta\} \tag{8}$$

for distinct $\xi, \eta \in S_*$, and $X \times \{\theta\}$ is closed in (F_ξ, μ_ξ), by (PR2). Moreover, it follows from (D1) that we may apply Lemma 2.3 for the family $\{(F_\xi, \mu_\xi) \mid \xi \in S_*\}$. Let μ be the metric on $\bigcup_{\xi \in S_*} F_\xi = X_*$ obtained by that lemma. Then

$$\hat{u}_\beta|_{X_*} \in \mathrm{Iso}(X_*, \mu) \quad (u \in G) \tag{9}$$

(see (D2) and the last claim in Lemma 2.3) and

$$\mu \text{ respects } d \quad \text{and} \quad \mu \leqslant 22 \tag{10}$$

(by (D1)). What is more,

($\star$) if $g \in \mathrm{Iso}(X_*, \mu)$ is such that $g(A) = A$ for any

$$A \in \mathcal{S} := \{X^n \times \{\xi\} \mid n \geqslant 1,\ \xi \in D_\beta\} \cup \{\{\xi\} \mid \xi \in S_*\},$$

then $g = \hat{u}_\beta|_{X_*}$ for some $u \in G$.

Let us briefly show ($\star$). If g is as specified there then $g(F_\xi) = F_\xi$ for any $\xi \in S_*$. So we infer from (D3) that $g = \widehat{(u_\xi)}_\beta$ on F_ξ for some $u_\xi \in \mathrm{Iso}(X, d)$ with

$$(u_\xi)^{ⓝ}(\kappa_n(\xi)) \in B_\xi, \tag{11}$$

where $n = n(\xi)$. We conclude from (8) that $u := u_\xi$ is independent of the choice of $\xi \in S_*$. Consequently, $g = \hat{u}_\beta|_{X_*}$. To end the proof of ($\star$), it remains to check that $u \in G$. Since the κ_n are surjective, (11) yields that $(u(z_1), \ldots, u(z_n))$ belongs to the closure (in X^n) of $\{(f(z_1), \ldots, f(z_n)) \mid f \in G\}$ for any $n \geqslant 2$ and $z_1, \ldots, z_n \in Z$. But this, combined with the fact that the function $\mathrm{Iso}(X, d) \ni f \mapsto f|_Z \in X^Z$ is an embedding (when X^Z is equipped with the pointwise convergence topology), yields that u belongs to the closure of G in $\mathrm{Iso}(X, d)$. But G is a closed subgroup, so we are done.

By Lemma 3.1, there is a metric

$$\lambda : S_0 \times S_0 \to \{0\} \cup [1, 2] \tag{12}$$

for which $\mathrm{Iso}(S_0, \lambda) = \{\mathrm{id}_{S_0}\}$ (id_{S_0} is the identity map on S_0). Let $\mathcal{S}$ be as in ($\star$). Since $\mathrm{card}(\mathcal{S}) = \beta = \mathrm{card}(S_0)$, there is a one-to-one function $v : \mathcal{S} \to \{11, 12\}^{S_0}$. We define a metric ϱ on $X_* \sqcup S_0 = \widehat{X}_\beta \setminus \{\theta\}$ by the rules:

- $\varrho = \mu$ on $X_* \times X_*$.
- $\varrho = \lambda$ on $S_0 \times S_0$.
- $\varrho(\xi, \eta) = \varrho(\eta, \xi) = [v(A)](\eta)$ for $\xi \in X_*$ and $\eta \in S_0$, where $A \in \mathcal{S}$ is such that $\xi \in A$ (such A is unique).

That ϱ is indeed a metric follows from (12), (10), axiom (PR2) for μ, and the fact that for any $\eta \in S_0$, $\varrho(\cdot, \eta)$ is constant on each member of $\mathcal{S}$. Finally, we extend the metric ϱ to $\widehat{X}_\beta$ by putting $\varrho(\xi, \theta) = 22$ for $\xi \in X_*$ and $\varrho(\xi, \theta) = 23$ for $\xi \in S_0$. Direct calculations show that ϱ is indeed a metric on $\widehat{X}_\beta$ and that ϱ respects d. It remains to check that (6) is a well-defined surjection (compare the proof of Proposition 2.2). We infer from (9) and the fact that $\varrho(\cdot, \eta)$ is constant on each member of $\mathcal{S}$ for any $\eta \in S_0 \cup \{\theta\}$ that the function (6) is well-defined. Now let $g \in \mathrm{Iso}(\widehat{X}_\beta, \varrho)$. Since θ is a unique point $\omega \in \widehat{X}_\beta$ such that $\mathrm{card}(\{\xi \in \widehat{X}_\beta \mid \varrho(\xi, \omega) = 23\}) > 1$, we obtain $g(\theta) = \theta$. Consequently, $g(X_*) = X_*$ and $g(S_0) = S_0$. The latter yields that

$$g|_{S_0} = \mathrm{id}_{S_0} \tag{13}$$

(because ϱ extends λ). Now if $\xi, \eta \in X_*$ are arbitrary, the injectivity of v and the definition of ϱ imply that $\varrho(\xi, \cdot) = \varrho(\eta, \cdot)$ on S_0 if and only if ξ and η belong to a common member of $\mathcal{S}$. But this, combined with (13), allows us to conclude that $g(A) = A$ for any $A \in \mathcal{S}$. Now an application of ($\star$) (recall that ϱ extends μ) provides us the existence of $u \in G$ for which $g = \hat{u}_\beta$ on X_*. Since $g(\xi) = \xi$ for $\xi \notin X_*$, we see that $g = \hat{u}_\beta$, which finishes the proof. $\square$

Remark 3.3. Under the notation and the assumptions of Theorem 3.2, if $M \geqslant 1$ is such that $d \leqslant M$ and $\varepsilon > 0$ is arbitrary, the metric ϱ appearing in the assertion of that theorem may be chosen so that $\varrho \leqslant M + \varepsilon$. Indeed, the above proof provides us the existence of a bounded metric ϱ, say $\varrho \leqslant C$, where $M < C < \infty$. Now if ε is small enough (that is, if $M + \varepsilon \leqslant C$), it suffices to replace ϱ by $\omega \circ \varrho$, where $\omega : [0, C] \to [0, M + \varepsilon]$ is affine on $[0, M]$ and $[M, C]$, and $\omega(0) = 0$, $\omega(M) = M$, and $\omega(C) = M + \varepsilon$.

Let (X, d) be a nonempty metric space and (Y, ϱ) denote the completion of (X, d). Since every isometry of (X, d) extends to a unique isometry of (Y, ϱ), the topological group $\mathrm{Iso}(X, d)$ may naturally be identified with the subgroup $\{u \in \mathrm{Iso}(Y, \varrho) \mid u(X) = X\}$ of $\mathrm{Iso}(Y, \varrho)$. If we follow this idea, Theorem 3.2 may be strengthened as follows:

Proposition 3.4. *Let (X, d) be a nonempty bounded metric space and (Y, ϱ) denote its completion. Let β be an infinite cardinal not less than $w(X)$. Further, let G be a closed subgroup of* $\mathrm{Iso}(X, d) \subset \mathrm{Iso}(Y, \varrho)$ *and $\bar{G}$ denote its closure in* $\mathrm{Iso}(Y, \varrho)$. *There are a complete metric λ on $\widehat{Y}_\beta$ respecting ϱ and a dense set $X_\beta \subset \widehat{Y}_\beta$ such that $(\widehat{Y}_\beta \setminus X_\beta, \lambda)$ is isometric to $(Y \setminus X, \varrho)$, and the function*

$$\bar{G} \ni u \mapsto \hat{u}_\beta \in \mathrm{Iso}(\widehat{Y}_\beta, \lambda) \tag{14}$$

is a well-defined isomorphism of topological groups which transforms G onto the group of all $u \in \mathrm{Iso}(\widehat{Y}_\beta, \lambda)$ with $u(X_\beta) = X_\beta$.

Proof. Fix $\theta \in D_\beta$ and put $X_\beta = \widehat{Y}_\beta \setminus [(Y \setminus X) \times \{\theta\}]$. By Theorem 3.2, there is a metric λ on $\widehat{Y}_\beta$ which respects ϱ and for which (14) is a well-defined isomorphism. Note that then λ is complete (see (AX4)), X_β is dense in $\widehat{Y}_\beta$, and $(\widehat{Y}_\beta \setminus X_\beta, \lambda)$ is isometric to $(Y \setminus X, \varrho)$ (since λ respects ϱ). Finally, if $u \in \bar{G}$ then $\hat{u}_\beta(X_\beta) = X_\beta$ if and only if $u(X) = X$ (which follows from the formulas for $\hat{u}_\beta$ and X_β). Equivalently, $\hat{u}_\beta(X_\beta) = X_\beta$ if and only if $u \in \mathrm{Iso}(X, d) \cap \bar{G} = G$, by the closedness of G in $\mathrm{Iso}(X, d)$. This shows the last claim of the theorem. $\square$

This proposition will be applied in Section 6, which is devoted to isometry groups of completely metrizable metric spaces.

To complete the proof of Theorem 3.2, we need to show Lemma 3.1. But the latter result immediately follows from the following much stronger result:

Proposition 3.5. *Let a and b be two reals such that*

$$0 < a < b \leqslant 2a. \tag{15}$$

For every set X having more than 5 points there is a metric $d : X \times X \to \{0, a, b\}$ such that

$$\mathrm{Iso}(X, d) = \{\mathrm{id}_X\}. \tag{16}$$

Proof. First of all, observe that any function $d : X \times X \to \{0, a, b\}$ which is symmetric and vanishes precisely on the diagonal of X is automatically a complete metric, which follows from (15). So we only need to take care of (16). For the same reason, we may (and do) assume, with no loss of generality, that $a = 1$ and $b = 2$. We shall make use of transfinite induction with respect to $\beta = \mathrm{card}(X) > 5$. Everywhere below in this proof, for $x \in X$, by $S(x)$ we denote the set of all $y \in X$ with $d(x, y) = 1$. Since we have to define a metric taking values in $\{0, 1, 2\}$, it is readily seen that it suffices to describe the sets $S(x)$ $(x \in X)$.

First assume $\beta = n \geqslant 6$ is finite. We may assume that $X = \{1, \dots, n\}$. Our metric d is defined by the following rules: $S(1) = \{2\}$, $S(2) = \{1, 3, 4, 5\}$, $S(3) = \{2, 4\}$, $S(4) = \{2, 3, 5\}$, $S(5) = \{2, 4, 6\}$, $S(n) = \{n - 1\}$, and $S(j) = \{j - 1, j + 1\}$ if $5 < j < n$. Take $g \in \mathrm{Iso}(X, d)$ and observe that:

- $g(2) = 2$, since 2 is the only point $j \in X$ such that $\mathrm{card}(S(j)) = 4$.
- $g(1) = 1$, because 1 is the unique point $j \in X$ for which $S(j) = \{2\}$.
- $g(3) = 3$, since 3 is the only point $j \in X$ such that $\mathrm{card}(S(j)) = 2$ and $2 \in S(j)$.
- $g(4) = 4$, because 4 is the unique point $j \in X$ for which $2 \neq j \in S(3)$.
- $g(5) = 5$, since 5 is the only point $j \in X$ such that $j \in S(4) \setminus \{2, 3\}$.

Now it is easy to check, using induction, that $g(j) = j$ for $j = 6, \dots, n$.

When $\beta = \aleph_0$, we may assume $X = \mathbb{N}$. Define a metric $d : \mathbb{N} \times \mathbb{N} \to \{0, 1, 2\}$ by $d(n, m) = \min(|m - n|, 2)$. It is left to the reader that $\mathrm{Iso}(\mathbb{N}, d) = \{\mathrm{id}_{\mathbb{N}}\}$ (use induction to show that $g(n) = n$ for any $n \in \mathbb{N}$ and $g \in \mathrm{Iso}(\mathbb{N}, d)$). Below we assume that $\beta > \aleph_0$ is such that for every infinite $\alpha < \beta$ the proposition holds for an arbitrary set X of cardinality α. For simplicity, for any uncountable cardinal γ we denote by I_γ the set of all cardinals α for which $\aleph_0 \leqslant \alpha < \gamma$. To get the assertion, we consider three cases.

First assume β is not limit; that is, β is the immediate successor of an infinite cardinal α. We may assume that X is the union of three pairwise disjoint sets X', $X' \times Y$, and $\{a\}$, where $\mathrm{card}(X') = \alpha$ and $\mathrm{card}(Y) = \beta$. It follows from the transfinite induction hypothesis that there exists a metric $d' : X' \times X' \to \{0, 1, 2\}$ such that $\mathrm{Iso}(X', d') = \{\mathrm{id}_{X'}\}$. Since $\beta \leqslant 2^\alpha$, there exists a one-to-one function

$\mu : X' \times Y \to \{1, 2\}^{X'}$ such that

$$[\mu(x, y)](x) = 1 \quad (x \in X', y \in Y) \tag{17}$$

(such a function μ may easily be constructed by transfinite induction with respect to an initial well-order on $X' \times Y$). We now define a metric d on X (with values in $\{0, 1, 2\}$) by the rules:

(d1) $d = d'$ on $X' \times X'$.

(d2) $d((x, y), (x', y')) = 1$ if (x, y) and (x', y') are distinct elements of $X' \times Y$.

(d3) $d((x, y), x') = [\mu(x, y)](x')$ if $x, x' \in X'$ and $y \in Y$.

(d4) $d(x, a) = 1$ and $d((x, y), a) = 2$ for any $x \in X'$ and $y \in Y$.

Observe that $S(x) \supset \{x\} \times Y$ and $S(x, y) \supset (X' \times Y) \setminus \{(x, y)\}$ for any $x \in X'$ and $y \in Y$ (thanks to (17) and (d2)–(d3)), and

$$S(a) = X' \tag{18}$$

(by (d4)). We infer from these facts that a is a unique point $x \in X$ such that $\mathrm{card}(S(x)) = \alpha$. Consequently, if $g \in \mathrm{Iso}(X, d)$ then $g(a) = a$ and $g(X') = X'$ (because of (18)). So $g|_{X'} \in \mathrm{Iso}(X', d')$ and therefore $g(x) = x$ for any $x \in X'$. Finally, if $x, x' \in X'$ and $y \in Y$ are arbitrary then $g(x, y) \notin X'$ and $d(g(x, y), x') = d((x, y), x')$, which yields that $\mu(g(x, y)) = \mu(x, y)$. So $g(x, y) = (x, y)$ (since μ is one-to-one) and we are done.

Now we assume that β is limit and $\mathrm{card}(I_\beta) < \beta$. For simplicity, put $I = I_\beta$. Let $\{X_\alpha\}_{\alpha \in I}$ be a family of pairwise disjoint sets such that

$$X_\alpha \cap I = \varnothing \quad \text{and} \quad \mathrm{card}(X_\alpha) = \alpha < \beta \quad (\alpha \in I). \tag{19}$$

Note that the set $X_* = \bigsqcup_{\alpha \in I} X_\alpha$ is of cardinality β and therefore we may assume $X = \{\omega\} \sqcup I \sqcup X_*$ (recall that this notation means that $\omega \notin I \cup X_*$). It follows from the transfinite induction hypothesis that there are metrics $d_I : I \times I \to \{0, 1, 2\}$ and $d_\alpha \mid X_\alpha \times X_\alpha \to \{0, 1, 2\}$ $(\alpha \in I)$ for which the groups $\mathrm{Iso}(I, d_I)$ and $\mathrm{Iso}(X_\alpha, d_\alpha)$ are trivial. We define a metric d on X as follows:

(d1′) $d = d_I$ on $I \times I$ and $d = d_\alpha$ on $X_\alpha \times X_\alpha$ for any $\alpha \in I$.

(d2′) $d(x, y) = 1$ if x and y belong to different members of the collection $\{X_\alpha\}_{\alpha \in I}$.

(d3′) $d(\alpha, x) = 2$ for $x \in X_\alpha$ and $d(\alpha, x) = 1$ for $x \in X_* \setminus X_\alpha$ $(\alpha \in I)$.

(d4′) $d(\alpha, \omega) = 1$ and $d(x, \omega) = 2$ for any $\alpha \in I$ and $x \in X_*$.

Observe that for any $\alpha \in I$ and $x \in X_\alpha$, we have $S(\alpha) \supset X_* \setminus X_\alpha$ by (d3′), and $S(x) \supset X_* \setminus X_\alpha$, by (d2′). At the same time, $S(\omega) = I$, by (d4′), and hence ω is the unique point $x \in X$ such that $\mathrm{card}(S(x)) < \beta$ (see (19)). Consequently, if $g \in \mathrm{Iso}(X, d)$ then $g(\omega) = \omega$, $g(I) = I$, and $g(X_*) = X_*$. Then $g|_I \in \mathrm{Iso}(I, d_I)$

(see (d1$'$)) and hence $g(\alpha) = \alpha$ for each $\alpha \in I$. Now use (d3$'$) to conclude that $g(X_\alpha) = X_\alpha$ for any $\alpha \in I$. So, according to (d1$'$), $g|_{X_\alpha} \in \mathrm{Iso}(X_\alpha, d_\alpha)$ for every $\alpha \in I$ and consequently $g(x) = x$ for all $x \in \bigcup_{\alpha \in I} X_\alpha = X_*$, and we are done.

Finally, assume β is limit and $\mathrm{card}(I_\beta) = \beta$. Then we may assume $X = I_\beta$. Since

$$\mathrm{card}(I_\alpha) \leqslant \alpha < \beta \tag{20}$$

whenever $\alpha \in X$, for every $\alpha \in X$ there is a cardinal $\gamma(\alpha) \in X$ such that

$$\mathrm{card}(\{\xi \mid \alpha < \xi \leqslant \gamma(\alpha)\}) = \alpha. \tag{21}$$

Now define a metric $d : X \times X \to \{0, 1, 2\}$ by the following rule: If $\aleph_0 \leqslant \alpha_1 < \alpha_2 < \beta$ then $d(\alpha_1, \alpha_2) = 1$ if and only if $\alpha_2 \leqslant \gamma(\alpha_1)$. It is easy to check that then $\mathrm{card}(S(\alpha)) = \alpha$ for any $\alpha \in X$ (thanks to (20) and (21)) and hence the identity map is a unique isometry on (X, d). □

4. Models for $\mathscr{G}_\delta$-complete groups

We begin this section with a useful characterization of $\mathscr{G}_\delta$-complete groups.

Proposition 4.1. *For a topological group G all conditions stated below are equivalent*:

- (I) *G is $\mathscr{G}_\delta$-complete.*
- (II) *G is isomorphic to a $\mathscr{G}_\delta$-closed subgroup of a complete topological group.*
- (III) *G is $\mathscr{G}_\delta$-closed in every topological group which contains G as a topological subgroup.*
- (IV) *Every net $\{x_\sigma\}_{\sigma \in \Sigma}$ of elements of G satisfying the following condition is convergent in G*:
  - (CC) *For every sequence $U_1, U_2, \ldots$ of neighborhoods of the neutral element of G there exist points $y, z \in G$ and a sequence $\sigma_1, \sigma_2, \dots \in \Sigma$ such that both $x_\sigma^{-1} y$ and $x_\sigma z^{-1}$ belong to U_n whenever $n \geqslant 1$ and $\sigma \geqslant \sigma_n$.*
- (V) *Every net $\{x_\sigma\}_{\sigma \in \Sigma}$ of elements of G satisfying the following condition is convergent in G*:
  - (CC$'$) *For every continuous left-invariant pseudometric d on G there are points $y, z \in G$ such that* $\lim_{\sigma \in \Sigma} d(x_\sigma, y) = \lim_{\sigma \in \Sigma} d(x_\sigma^{-1}, z^{-1}) = 0$.

Proof. Everywhere below, τ is the topology of G and e is its neutral element.

First assume G is a $\mathscr{G}_\delta$-closed subgroup of a complete group H. We want to show that G is $\mathscr{G}_\delta$-complete. Let $\{x_\sigma\}_{\sigma \in \Sigma} \subset G$ be a net which satisfies condition (C) with respect to the topology τ_δ. It then satisfies this condition with respect to τ as well. Since H is complete, there is $y \in H$ such that $\lim_{\sigma \in \Sigma} x_\sigma = y$. It suffices to check that y belongs to the $\mathscr{G}_\delta$-closure of G in H. Take a $\mathscr{G}_\delta$-subset A of H

containing y and write Ay^{-1} in the form $Ay^{-1} = \bigcap_{n=1}^{\infty} U_n$, where each U_n is open in H and contains e. Using the regularity of the space H, we may find a sequence $V_1, V_2, \dots$ of open (in H) neighborhoods of e such that the closure (in H) of V_n is contained in $V_{n-1} \cap U_n$ for each n, where $V_0 = H$. Then $F := \bigcap_{n=1}^{\infty} V_n$ is a closed $\mathscr{G}_\delta$-subset of H and $e \in F \subset Ay^{-1}$. It follows from our assumption about the net that there is $\sigma_0 \in \Sigma$ such that $x_{\sigma_0} x_\sigma^{-1} \in F$ for any $\sigma \geqslant \sigma_0$. We now infer from the closedness of F in H that $x_{\sigma_0} y^{-1} \in F$ as well, and consequently, $x_{\sigma_0} \in A$, which shows that y belongs to the $\mathscr{G}_\delta$-closure of G. This proves that (II) is followed by (I). Conversely, if G is a $\mathscr{G}_\delta$-complete subgroup of a topological group K and $\mathbb{O}$ denotes the topology of K then τ_δ coincides with the topology (on G) of a subspace inherited from $(K, \mathbb{O}_\delta)$. It now follows from the completeness of (G, τ_δ) that G is closed in $(K, \mathbb{O}_\delta)$ or, equivalently, that G is $\mathscr{G}_\delta$-closed in K, which proves that (III) is implied by (I). Since (II) obviously follows from (III), in this way we have shown that conditions (I), (II), and (III) are equivalent. We shall now show that (II) is equivalent to (IV) and then that (IV) is equivalent to (V).

If (II) is fulfilled then G is $\mathscr{G}_\delta$-closed in $\overline{G}$. Let $\{x_\sigma\}_{\sigma\in\Sigma}$ be a net of elements of G which satisfies condition (CC). Then it fulfills condition (C) as well and hence there is $w \in \overline{G}$ such that $\lim_{\sigma\in\Sigma} x_\sigma = w$. It suffices to check that $w \in G$ or, equivalently, that w belongs to the $\mathscr{G}_\delta$-closure of G. To this end, take any $\mathscr{G}_\delta$-subset A of $\overline{G}$ which contains w. Write $Aw^{-1} = \bigcap_{n=1}^{\infty} V_n$, where $V_1, V_2, \dots$ are open neighborhoods of e. For each $n \geqslant 1$ take a neighborhood U_n of e with $U_n = U_n^{-1}$ and $U_n \cdot U_n \subset V_n$. Now let y, z, and $\sigma_1, \sigma_2, \dots$ be as in (CC), applied for the sequence

$$U_1 \cap G, U_2 \cap G, \dots.$$

Fix for a moment $n \geqslant 1$. Choose $\sigma \geqslant \sigma_n$ such that $x_\sigma w^{-1} \in U_n$. Then $zw^{-1} = (x_\sigma z^{-1})^{-1}(x_\sigma w^{-1}) \subset U_n^{-1} \cdot U_n \subset V_n$. So $zw^{-1} \in \bigcap_{n=1}^{\infty} V_n = Aw^{-1}$, which implies that $w \in A$. Consequently, $A \cap G \neq \varnothing$ and we are done.

The converse implication goes similarly: When (IV) is satisfied, we show that G is $\mathscr{G}_\delta$-closed in $\overline{G}$. Let $w \in \overline{G}$ belong to the $\mathscr{G}_\delta$-closure of G. Then, of course, w is in the closure of G and thus there is a net $\{x_\sigma\}_{\sigma\in\Sigma} \subset G$, which converges to w. To prove that $w \in G$, it is enough to verify that (CC) is fulfilled. To this end, fix a sequence $U_1, U_2, \dots$ of neighborhoods of e and choose its open symmetric neighborhoods $V_1, V_2, \dots$ such that $V_n \cdot V_n \subset U_n$ $(n \geqslant 1)$. We conclude from the fact that w is in the $\mathscr{G}_\delta$-closure of G that there is $y \in G$ such that $y \in \bigcap_{n=1}^{\infty}(V_n w \cap w V_n)$. Fix $n \geqslant 1$. There is $\sigma_n \in \Sigma$ such that both $x_\sigma w^{-1}$ and $w^{-1} x_\sigma$ belong to V_n for $\sigma \geqslant \sigma_n$. Then, for such σ's, $x_\sigma y^{-1} = (x_\sigma w^{-1})(yw^{-1})^{-1} \subset V_n \cdot V_n^{-1} \subset U_n$ and $x_\sigma^{-1} y = (w^{-1} x_\sigma)^{-1}(w^{-1} y) \subset V_n^{-1} \cdot V_n \subset U_n$ as well. This shows that (CC) is satisfied for $z = y$, and we are done.

Point (V) is easily implied by (IV) (for a fixed continuous left-invariant pseudometric d and a net satisfying (CC$'$) apply (CC) for $U_n = \{x \in G \mid d(x, e) < 2^{-n}\}$).

The converse implication follows from the well-known fact that for an arbitrary sequence $U_1, U_2, \dots$ of neighborhoods of e there exists a left-invariant pseudometric d on G such that $\{x \in G \mid d(x, e) < 2^{-n}\} \subset U_n$ for every $n \geqslant 1$ (see, for example, the proof of the Kakutani–Birkhoff theorem on the metrizability of topological groups presented in [Berberian 1974, Theorem 6.3]; or use Markov's theorem [Arhangel'skii and Tkachenko 2008, Theorem 3.3.9] to deduce this property). □

Remark 4.2. The proof of Proposition 4.1 shows that points (IV) and (V) of that result may be weakened by assuming that every net satisfying condition (CC) or (CC$'$) with $z = y$ is convergent. However, to prove Theorem 1.4, we need (IV) in its present form.

Now we can give many examples of $\mathscr{G}_\delta$-complete groups. We inform that by the Cartesian product of a family $\{G_s\}_{s\in S}$ of topological groups we mean the "full" Cartesian product $\prod_{s\in S} G_s$ of them and by the direct product of this family we mean the topological subgroup $\bigoplus_{s\in S} G_s$ of $\prod_{s\in S} G_s$ consisting of all its finitely supported elements.

Proposition 4.3. *Each of the following topological groups is $\mathscr{G}_\delta$-complete*:

(a) *A $\mathscr{G}_\delta$-closed subgroup of a $\mathscr{G}_\delta$-complete group. A complete group.*

(b) *The Cartesian as well as the direct product of arbitrary collection of $\mathscr{G}_\delta$-complete groups.*

(c) *A topological group which is the countable union of its subgroups each of which is $\mathscr{G}_\delta$-complete.*

(d) *A topological group which is σ-compact as a topological space. In particular, all countable topological groups are $\mathscr{G}_\delta$-complete.*

(e) *A topological group in which singletons are $\mathscr{G}_\delta$. In particular, metrizable groups are $\mathscr{G}_\delta$-complete.*

(f) $G = \operatorname{Iso}(X, d)$ *for an arbitrary metric space* (X, d). *Moreover*, $w(G) \leqslant w(X)$, *and G is complete provided* (X, d) *is complete.*

Proof. In each point we invoke Proposition 4.1.

To prove point (a), use the equivalence between conditions (I) and (III) in Proposition 4.1. We turn to (b). Let $\{G_s\}_{s\in S}$ be a nonempty collection of $\mathscr{G}_\delta$-complete groups and let $G = \prod_{s\in S} G_s$. Let $x_\sigma = (x_\sigma^{(s)})_{s\in S} \in G$ ($\sigma \in \Sigma$) be a net satisfying condition (CC). It remains to check that for any $t \in S$, the net $\{x_\sigma^{(t)}\}_{\sigma\in\Sigma} \subset G_t$ satisfies condition (CC) as well (because then it will be convergent), which is immediate: If $V_1, V_2, \dots$ is a sequence of neighborhoods of the neutral element of G_t, apply (CC) for the sequence $U_1, U_2, \dots$ with $U_j := \{(x^{(s)})_{s\in S} \in G \mid x^{(t)} \in V_j\}$ ($j \geqslant 1$) to obtain two points $y, z \in G$ and then use their t-coordinates. Now to prove that $H := \bigoplus_{s\in S} G_s$ is also $\mathscr{G}_\delta$-complete, it suffices to check that it is

$\mathcal{G}_\delta$-closed in G (by (a)). But if $y = (y_s)_{s\in S} \in G \setminus H$, there is a countable (infinite) set $S' \subset S$ such that $y_s \neq e_s$ for any $s \in S'$, where e_s is the neutral element of G_s. Then the set $A := \{(z_s)_{s\in S} \in G \mid z_s \neq e_s \text{ for all } s \in S'\}$ is a $\mathcal{G}_\delta$-subset of G which contains y and is disjoint from H, which finishes the proof of (b).

Since the proofs of points (c) and (d) are similar, we shall show only (c). Let $G = \bigcup_{n=1}^{\infty} G_n$, where G_n is $\mathcal{G}_\delta$-complete for any n. Let $y \in \overline{G} \setminus G$. Then $y \notin G_n$ and G_n is $\mathcal{G}_\delta$-closed in $\overline{G}$. Consequently, there are $\mathcal{G}_\delta$-subsets $A_1, A_2, \ldots$ of $\overline{G}$ containing y such that $A_n \cap G_n = \varnothing$. Then $A := \bigcap_{n=1}^{\infty} A_n$ is also a $\mathcal{G}_\delta$-subset of $\overline{G}$ containing y, and $A \cap G = \varnothing$. This shows that y is not in the $\mathcal{G}_\delta$-closure of G and we are done.

Further, if all singletons are $\mathcal{G}_\delta$ in G then τ_δ is discrete and hence G is $\mathcal{G}_\delta$-complete. This proves (e).

Finally, we turn to (f). The second and the third claims of (f) are well-known, but for the sake of completeness we shall prove them too. Let (X, d) be a metric space and $G = \mathrm{Iso}(X, d)$. Let $\{u_\sigma\}_{\sigma\in\Sigma} \subset G$ be a net satisfying condition (CC$'$). Fix $x \in X$ and put $\varrho : G \times G \ni (u, v) \mapsto d(u(x), v(x)) \in [0, \infty)$. Observe that ϱ is a left-invariant continuous pseudometric on G. It follows from (CC$'$) that there are $f, g \in G$ such that $\lim_{\sigma\in\Sigma} d(u_\sigma(x), f(x)) = \lim_{\sigma\in\Sigma} d(u_\sigma^{-1}(x), g(x)) = 0$. We conclude that both the nets $\{u_\sigma(x)\}_{\sigma\in\Sigma}$ and $\{u_\sigma^{-1}(x)\}_{\sigma\in\Sigma}$ converge in X. So we may define $u, v : X \to X$ by $u(x) = \lim_{\sigma\in\Sigma} u_\sigma(x)$ and $v(x) = \lim_{\sigma\in\Sigma} u_\sigma^{-1}(x)$ $(x \in X)$. It is readily seen that both u and v are isometric. What is more, a standard argument proves that $u \circ v = v \circ u = \mathrm{id}_X$ and hence $u \in G$ and $\lim_{\sigma\in\Sigma} u_\sigma = u$. So G is $\mathcal{G}_\delta$-complete. Furthermore, if D is a dense subset of X such that $\mathrm{card}(D) = w(X)$ then the function $G \ni g \mapsto g|_D \in X^D$ is a topological embedding (when X^D is equipped with the pointwise convergence topology) and therefore $w(G) \leqslant w(X^D) \leqslant w(X)$. Finally, if (X, d) is in addition complete and $\{u_\sigma\}_{\sigma\in\Sigma} \subset G$ is a net satisfying (C), similar argument to that above shows that then for any $x \in X$ the nets $\{u_\sigma(x)\}_{\sigma\in\Sigma}$ and $\{u_\sigma^{-1}(x)\}_{\sigma\in\Sigma}$ are fundamental in (X, d) and hence converge. It now follows from the previous part of the proof that $\{u_\sigma\}_{\sigma\in\Sigma}$ is convergent in G, which finishes the proof. □

For the purpose of the next result, recall that a topological space X is *Dieudonné-complete* if and only if there is a complete uniformity on X inducing the topology of X (see, for example, [Engelking 1989, Chapter 8]). Accordingly, a topological group is *Dieudonné-complete* if and only if it is Dieudonné-complete as a topological space ([Arhangel'skii and Tkachenko 2008]).

Corollary 4.4. *For every metric space* (X, d) *the topological group* $\mathrm{Iso}(X, d)$ *is Dieudonné-complete.*

Proof. By Proposition 4.3, $\mathrm{Iso}(X, d)$ is $\mathcal{G}_\delta$-complete and hence it is $\mathcal{G}_\delta$-closed in $\overline{G}$ thanks to Proposition 4.1. Consequently, $\mathrm{Iso}(X, d)$ is Dieudonné-complete (since

$\overline{G}$ is such and $\mathscr{G}_\delta$-closed subsets of Dieudonné-complete spaces are Dieudonné-complete as well — see [Dieudonné 1939] or Problem 8.5.13(f) on page 465 in [Engelking 1989]; see also the proof of Proposition 6.5.2 on page 366 in [Arhangel'skii and Tkachenko 2008]). □

The above result gives a full answer to the question of when the isometry group of a metric space is Dieudonné-complete, posed by Arhangel'skii and Tkachenko [2008, Open Problem 3.5.4 on page 181].

Example 4.5. As we announced in the introductory part, not every topological group is absolutely $\mathscr{G}_\delta$-closed. Let us briefly justify our claim. Let S be an uncountable set and for each $s \in S$ let G_s be a nontrivial complete group with the neutral element e_s. Then $G := \prod_{s\in S} G_s$ is a complete group as well and

$$G_0 = \{(x_s)_{s\in S} \in G \mid \operatorname{card}(\{s \in S \mid x_s \neq e_s\}) \leqslant \aleph_0\}$$

is a proper subgroup of G which is $\mathscr{G}_\delta$-dense in G (and thus G_0 is not $\mathscr{G}_\delta$-closed in G). Indeed, if $y = (y_s)_{s\in S} \in G$ and A is a $\mathscr{G}_\delta$-subset of G containing y then there is a countable set $S_0 \subset S$ such that $\{(x_s)_{s\in S} \in G \mid x_s = y_s \text{ for all } s \in S_0\} \subset A$; then, $z \in G_0 \cap A$, where $z_s = y_s$ for $s \in S_0$ and $z_s = e_s$ otherwise.

We are almost ready to prove Theorem 1.4. For the purpose of its proof and the nearest result, let us introduce auxiliary notations and terminology. Whenever d and d' are two bounded pseudometrics on a common nonempty set X, we put

$$\|d - d'\|_\infty := \sup_{x,y\in X} |d(x, y) - d'(x, y)|.$$

Further, the relation $R := \{(x, y) \in X \times X \mid d(x, y) = 0\}$ is an equivalence on X. Let $\pi : X \to X/R$ be the canonical projection. The function $D : (X/R) \times (X/R) \ni (\pi(x), \pi(y)) \mapsto d(x, y) \in [0, \infty)$ is a well-defined metric on X/R. We call a triple $(Y, \varrho; \xi)$ a *metric space associated with* (X, d) if (Y, ϱ) is a metric space isometric to $(X/R, D)$ and ξ is a function of X onto Y such that there is an isometry $g : (X/R, D) \to (Y, \varrho)$ for which $\xi = g \circ \pi$. Observe that then $\varrho(\xi(x), \xi(y)) = d(x, y)$ for any $x, y \in X$.

With use of the following result we shall take care of condition $w(X) = w(G)$ in Theorem 1.4(A):

Lemma 4.6. *Let G be a topological group and $\{\varrho_s\}_{s\in S}$ be a collection of bounded continuous left-invariant pseudometrics on G. For each $s \in S$, let $(X_s, d_s; \pi_s)$ be a metric space associated with (G, ϱ_s) chosen so that the sets X_s are pairwise disjoint. There exists a metric d on $X := \bigcup_{s\in S} X_s$ with the following properties*:

(DD1) $\frac{1}{2}\sqrt[3]{d_s(x, y)} \leqslant d(x, y) \leqslant \sqrt[3]{d_s(x, y)}$ *for any $x, y \in X_s$ and $s \in S$.*

(DD2) $d(\pi_s(a), \pi_t(a)) \leqslant \sqrt[3]{\|\varrho_s - \varrho_t\|_\infty}$ *for any $a \in G$ and $s, t \in S$.*

(DD3) *Each of the sets* X_s $(s \in S)$ *is closed in* (X, d).

(DD4) $d(\pi_s(ag), \pi_t(ah)) = d(\pi_s(g), \pi_t(h))$ *for any* $a, g, h \in G$ *and* $s, t \in S$.

Proof. To simplify arguments, for each $x \in X$ denote by $\kappa(x)$ the unique index $s \in S$ such that $x \in X_s$. Define a function $v : X \times X \to [0, \infty)$ as follows:

$$v(x, y) = \|\varrho_{\kappa(x)} - \varrho_{\kappa(y)}\|_\infty + \inf\{d_{\kappa(x)}(x, \pi_{\kappa(x)}(g)) + d_{\kappa(y)}(\pi_{\kappa(y)}(g), y) \mid g \in G\}.$$

Observe that:

- $v(x, x) = 0$ and $v(y, x) = v(x, y)$ for any $x, y \in X$.
- $v(x, y) = d_s(x, y)$ for any $x, y \in X_s$ and $s \in S$.
- $v(\pi_s(g), \pi_t(g)) = \|\varrho_s - \varrho_t\|_\infty$ whenever $s, t \in S$ and $g \in G$.
- $v(\pi_s(g), \pi_t(h)) \geqslant \|\varrho_s - \varrho_t\|_\infty$ for all $s, t \in S$ and $g, h \in G$.
- $v(\pi_s(ag), \pi_t(ah)) = v(\pi_s(g), \pi_t(h))$ for any $a, g, h \in G$ and $s, t \in S$.

Let us now show that for any $x_0, x_1, x_2, x_3 \in X$ and each $\varepsilon > 0$

$$\max_{j=1,2,3} v(x_{j-1}, x_j) < \varepsilon \implies v(x_0, x_3) < 8\varepsilon. \tag{22}$$

Assume $v(x_{j-1}, x_j) < \varepsilon$ $(j = 1, 2, 3)$. This means that there are $a_1, a_2, a_3 \in G$ for which

$$\|\varrho_{\kappa(x_{j-1})} - \varrho_{\kappa(x_j)}\|_\infty + d_{\kappa(x_{j-1})}(x_{j-1}, \pi_{\kappa(x_{j-1})}(a_j)) + d_{\kappa(x_j)}(\pi_{\kappa(x_j)}(a_j), x_j) < \varepsilon. \tag{23}$$

In particular, $\|\varrho_{\kappa(x_{j-1})} - \varrho_{\kappa(x_j)}\|_\infty < \varepsilon$ for $j = 1, 2, 3$ and thus

$$\|\varrho_{\kappa(x_0)} - \varrho_{\kappa(x_3)}\|_\infty < 3\varepsilon. \tag{24}$$

For simplicity, for $j \in \{0, 1, 2, 3\}$ put $s_j = \kappa(x_j)$ and take $b_j \in G$ such that $\pi_{s_j}(b_j) = x_j$. Recall that $d_s(\pi_s(g), \pi_s(h)) = \varrho_s(g, h)$ for any $s \in S$ and $g, h \in G$. Therefore we have

$$\begin{aligned}\varrho_{s_j}(b_{j-1}, b_j) &\leqslant \varrho_{s_j}(b_{j-1}, a_j) + \varrho_{s_j}(a_j, b_j)\\ &\leqslant \|\varrho_{s_{j-1}} - \varrho_{s_j}\|_\infty + \varrho_{s_{j-1}}(b_{j-1}, a_j) + \varrho_{s_j}(a_j, b_j) < \varepsilon\end{aligned}$$

(by (23)) and consequently

$$\begin{aligned}\varrho_{s_2}(b_0, b_2) &\leqslant \varrho_{s_2}(b_0, b_1) + \varrho_{s_2}(b_1, b_2)\\ &\leqslant \|\varrho_{s_2} - \varrho_{s_1}\|_\infty + \varrho_{s_1}(b_0, b_1) + \varrho_{s_2}(b_1, b_2) < 3\varepsilon.\end{aligned}$$

Similarly,

$$\begin{aligned}\varrho_{s_3}(b_0, b_3) &\leqslant \varrho_{s_3}(b_0, b_2) + \varrho_{s_3}(b_2, b_3)\\ &\leqslant \|\varrho_{s_3} - \varrho_{s_2}\|_\infty + \varrho_{s_2}(b_0, b_2) + \varrho_{s_3}(b_2, b_3) < 5\varepsilon.\end{aligned} \tag{25}$$

Finally, by (24) and (25) we obtain

$$\begin{aligned} v(x_0, x_3) &\leqslant \|\varrho_{s_0} - \varrho_{s_3}\|_\infty + d_{s_0}(\pi_{s_0}(b_0), \pi_{s_0}(b_0)) + d_{s_3}(\pi_{s_3}(b_0), \pi_{s_3}(b_3)) \\ &< 3\varepsilon + \varrho_{s_3}(b_0, b_3) \\ &< 8\varepsilon, \end{aligned}$$

and the proof of (22) is complete. Now let $f : X \times X \to [0, \infty)$ be given by $f(x, y) = \sqrt[3]{v(x, y)}$. Below we collect all properties established for v and translated to the case of the function f:

(F1) $f(x, x) = 0$ and $f(x, y) = f(y, x) > 0$ for any distinct points x and y of X.

(F2) If $\varepsilon > 0$ and $\{f(x, y), f(y, z), f(z, w)\} \subset [0, \varepsilon]$ for some $x, y, z, w \in X$ then $f(x, w) \leqslant 2\varepsilon$, thanks to (22).

(F3) $f(x, y) = \sqrt[3]{d_s(x, y)}$ and $f(\pi_s(g), \pi_t(g)) = \sqrt[3]{\|\varrho_s - \varrho_t\|_\infty}$ whenever $x, y \in X_s$, $g \in G$, and $s, t \in S$.

(F4) $f(\pi_s(g), \pi_t(h)) \geqslant \sqrt[3]{\|\varrho_s - \varrho_t\|_\infty}$ for all $g, h \in G$ and $s, t \in S$.

(F5) $f(\pi_s(ag), \pi_t(ah)) = f(\pi_s(g), \pi_t(h))$ for any $a, g, h \in G$ and $s, t \in S$.

Finally, we define $d : X \times X \to [0, \infty)$ as follows:

$$d(x, y) = \inf\left\{\sum_{j=1}^{n} f(z_{j-1}, z_j) \,\middle|\, n \geqslant 1, z_0, \dots, z_n \in X, z_0 = x, z_n = y\right\}.$$

Lemma 6.2 of [Berberian 1974] asserts that for any function $f : X \times X \to [0, \infty)$ satisfying conditions (F1)–(F2) the function d defined above is a metric on X such that

$$\tfrac{1}{2} f(x, y) \leqslant d(x, y) \leqslant f(x, y) \quad (x, y \in X). \tag{26}$$

It follows from (F5) and the formula of d that (DD4) is fulfilled, while (DD1) and (DD2) may easily be deduced from (F3) and (26). Finally, (DD3) is a consequence of (F4) and (26). □

Proof of Theorem 1.4. The implications (A1) ⇒ (A2) and (B1) ⇒ (B2) follow from Proposition 4.3. It remains to show the converse implications.

First assume that G is complete (in this case the proof is much shorter). By a well-known result (see, for example, [Uspenskij 2008, Theorem 2.1]), there is a bounded metric space (Y, ϱ) such that $w(G) = w(Y)$ and G is isomorphic to a subgroup H of $\mathrm{Iso}(Y, \varrho)$. Since $\mathrm{Iso}(Y, \varrho)$ is naturally isomorphic to a subgroup of $\mathrm{Iso}(\bar{Y}, \bar{\varrho})$, where $(\bar{Y}, \bar{\varrho})$ is the completion of (Y, ϱ), we may assume that (Y, ϱ) is a complete metric space. Since G is complete, H is a closed subgroup of (Y, ϱ). Now Theorem 3.2 implies that H is isomorphic to $\mathrm{Iso}(X, d)$, where $X = \widehat{Y}_\beta$ with $\beta = w(Y)$, and d is a metric which respects ϱ. Notice that d is complete (by (AX4)),

$w(X) = w(G)$ (because $\beta = w(Y) = w(G)$) and G is isomorphic to $\mathrm{Iso}(X, d)$ (being isomorphic to H). This proves the remainder of point (B).

We now turn to (A). Assume G is $\mathscr{G}_\delta$-complete. Thanks to Theorem 3.2, it suffices to show that G is isomorphic to a closed subgroup of $\mathrm{Iso}(X, d)$ for a metric space (X, d) of topological weight equal to $w(G)$ (see the previous part of the proof). We shall do this employing Lemma 4.6 and improving a classical argument, presented, for example, in the first proof of [Uspenskij 2008, Theorem 2.1]. (That proof shows that every topological group is isomorphic to a topological subgroup of $\mathrm{Iso}(X, d)$ for some metric space (X, d). However, this fact is insufficient for us, not only because the topological weight of X is out of control. A much more difficult problem is to provide the closedness of the subgroup of $\mathrm{Iso}(X, d)$ isomorphic to a given $\mathscr{G}_\delta$-complete group.)

Let $\mathscr{B}$ be a base of open neighborhoods of the neutral element e of G such that $\mathrm{card}(\mathscr{B}) \leqslant w(G)$. Let S be the set of all finite and all infinite sequences of members of $\mathscr{B}$. For any $U \in \mathscr{B}$ there exists a continuous left-invariant pseudometric λ_U on G bounded by 1 such that

$$\{x \in G \mid \lambda_U(x, e) < 1\} \subset U. \tag{27}$$

We leave it as a simple exercise that the family $\{\lambda_U\}_{U \in \mathscr{B}}$ determines the topology of G. Now for any $s = (U_j)_{j=1}^N \in S$ (where N is finite or $N = \infty$) let

$$\varrho_s := \sum_{j=1}^{N} \frac{1}{2^j} \lambda_{U_j}. \tag{28}$$

Notice that ϱ_s is a continuous left-invariant pseudometric on G bounded by 1. What is more,

(T) the family $\{\varrho_s\}_{s \in S}$ determines the topology of G

(since $\varrho_s = \lambda_U$ for $s = (U, U, \dots) \in S$). Now we apply Lemma 4.6 for the family $\{\varrho_s\}_{s \in S}$. Let $(X_s, d_s; \pi_s)$ $(s \in S)$ and (X, d) be as stated there. That is, $(X_s, d_s; \pi_s)$ is a metric space associated with (G, ϱ_s), the sets X_s are pairwise disjoint, $X = \bigcup_{s \in S} X_s$, and d is a metric on X satisfying conditions (DD1)–(DD4). For each $s \in S$ let $\tilde{\varrho}_s : G \times G \to [0, \infty)$ be given by $\tilde{\varrho}_s(g, h) = d(\pi_s(g), \pi_s(h))$. It is clear that $\tilde{\varrho}_s$ is a pseudometric on G. What is more, it is left-invariant, thanks to (DD4), and moreover

$$\tfrac{1}{2}\sqrt[3]{\varrho_s} \leqslant \tilde{\varrho}_s \leqslant \sqrt[3]{\varrho_s} \quad (s \in S), \tag{29}$$

by (DD1). Consequently, each of the pseudometrics $\tilde{\varrho}_s$ is continuous and

(T′) the family $\{\tilde{\varrho}_s\}_{s \in S}$ determines the topology of G

(see (T)). We infer from the continuity of $\tilde{\varrho}_s$ that π_s, as a function of G into (X,d), is continuous as well. We claim that $w(X) \leqslant w(G)$. To see this, let S_f consists of all finite sequences of members of $\mathscr{B}$, and let D be a dense subset of G such that $\operatorname{card}(D) \leqslant w(G)$. Observe that $Z := \bigcup_{s\in S_f} \pi_s(D)$ has cardinality not exceeding $w(G)$. We will now show that Z is dense in X. First of all, note that $\pi_s(D)$ is dense in X_s, since π_s is continuous. In particular, the closure of Z contains all points of $\bigcup_{s\in S_f} X_s$. Fix $s \notin S_f$ and $a \in G$. Then s is of the form $s = (U_j)_{n=1}^{\infty} \in S$. Put $s_n := (U_j)_{j=1}^{n} \in S_f$ and observe that, by (DD2) and (28),

$$d(\pi_{s_n}(a), \pi_s(a)) \leqslant \sqrt[3]{\|\varrho_{s_n} - \varrho_s\|_\infty} \to 0 \quad (n \to \infty).$$

So, since $\pi_s(G) = X_s$, the above argument shows that Z is indeed dense in X.

It remains to check that G is isomorphic to a closed subgroup of $\operatorname{Iso}(X,d)$. For $g \in G$ let $u_g : X \to X$ be such that $u_g(\pi_s(x)) = \pi_s(gx)$ for any $s \in S$ and $x \in G$. Then $\Phi : G \ni g \mapsto u_g \in \operatorname{Iso}(X,d)$ is a well-defined (by (DD4)) group homomorphism as well as a topological embedding (thanks to (T′)). So it follows from Proposition 4.3(f) that $w(X) \geqslant w(G)$ and hence $w(X) = w(G)$. We shall check that $\Phi(G)$ is closed, which will finish the proof. Assume $\{x_\sigma\}_{\sigma\in\Sigma}$ is a net in G such that the net $\{u_{x_\sigma}\}_{\sigma\in\Sigma}$ converges in $\operatorname{Iso}(X,d)$ to some $u \in \operatorname{Iso}(X,d)$. It is enough to prove that the net $\{x_\sigma\}_{\sigma\in\Sigma}$ is convergent in G. Since G is $\mathscr{G}_\delta$-complete, actually it suffices to verify condition (CC) (see Proposition 4.1). To this end, let $V_1, V_2, \ldots$ be a sequence of neighborhoods of e. For any $n \geqslant 1$ choose $U_n \in \mathscr{B}$ for which $U_n \subset V_n$. Now for $s := (U_j)_{j=1}^{\infty} \in S$ we have

$$\lim_{\sigma\in\Sigma} d(\pi_s(x_\sigma), u(\pi_s(e))) = \lim_{\sigma\in\Sigma} d(u_{x_\sigma}(\pi_s(e)), u(\pi_s(e))) = 0,$$
$$\lim_{\sigma\in\Sigma} d(\pi_s(x_\sigma^{-1}), u^{-1}(\pi_s(e))) = \lim_{\sigma\in\Sigma} d(u_{x_\sigma}^{-1}(\pi_s(e)), u^{-1}(\pi_s(e))) = 0.$$

We infer from (DD3) and the above convergences that there are $y, z \in G$ such that $\lim_{\sigma\in\Sigma} d(\pi_s(x_\sigma), \pi_s(y)) = \lim_{\sigma\in\Sigma} d(\pi_s(x_\sigma^{-1}), \pi_s(z^{-1})) = 0$. But for any $a, b \in G$, $d(\pi_s(a), \pi_s(b)) = \tilde{\varrho}_s(a,b)$ and thus, thanks to (29),

$$\lim_{\sigma\in\Sigma} \varrho_s(x_\sigma, y) = \lim_{\sigma\in\Sigma} \varrho_s(x_\sigma^{-1}, z^{-1}) = 0.$$

For each $n \geqslant 1$ let $\sigma_n \in \Sigma$ be such that both the numbers $\varrho_s(x_\sigma, y) = \varrho_s(x_\sigma^{-1}y, e)$ and $\varrho_s(x_\sigma^{-1}, z^{-1}) = \varrho_s(x_\sigma z^{-1}, e)$ are less than 2^{-n} for any $\sigma \geqslant \sigma_n$. We deduce from the formula of s, (28), and (27) that $\{x_\sigma^{-1}y, x_\sigma z^{-1}\} \subset U_n (\subset V_n)$ for all $\sigma \geqslant \sigma_n$. $\square$

The above proof provides the existence of a metric space (namely, $\widehat{X}_\beta$) whose isometry group is isomorphic to a given $\mathscr{G}_\delta$-complete group G. This metric space is highly disconnected (since it contains a clopen discrete set whose cardinality is equal to the topological weight of the whole space). In the next section we shall improve Theorem 1.4 by showing that G is isomorphic to the isometry group of a contractible open set in a normed vector space of the same topological weight as G.

Corollary 4.7. (A) *Let G be a topological group and β be an infinite cardinal number. There exists a* (*complete*) *metric space* (X, d) *such that* $w(X) = \beta$ *and* $\mathrm{Iso}(X, d)$ *is isomorphic to G if and only if G is $\mathscr{G}_\delta$-complete* (*resp. complete*) *and* $\beta \geqslant w(G)$.

(B) *A topological group is isomorphic to the isometry group of some separable metric space if and only if it is second-countable.*

Proof. Both points (A) and (B) follow from Theorems 1.4 and 3.2, (AX4) and, respectively, points (f) and (e) of Proposition 4.3. □

We call a metrizable space X *zero-dimensional* if and only if X has a base consisting of clopen (that is, simultaneously open and closed) sets; X is *strongly zero-dimensional* if the covering dimension of X equals 0.

Corollary 4.8. *Let G be an infinite metrizable topological group and $\beta = w(G)$. There exists a compatible metric ϱ on $Y := \widehat{G}_\beta$ such that G is isomorphic to* $\mathrm{Iso}(Y, \varrho)$. *In particular*:

(A) *If G is discrete*, *there is a complete compatible* (*possibly nonleft-invariant*) *metric d on G such that G is isomorphic to* $\mathrm{Iso}(G, d)$.

(B) *If G is countable and nondiscrete*, *there is a compatible metric d on $F := \mathbb{Q} \sqcup \mathbb{Z}$ such that G is isomorphic to* $\mathrm{Iso}(F, d)$.

(C) *If G is totally disconnected* (*zero-dimensional*; *strongly zero-dimensional*) *then there is a metric space* (X, d) *such that X is totally disconnected* (*resp. zero-dimensional*; *strongly zero-dimensional*) *as well*, $w(X) = w(G)$, *and* $\mathrm{Iso}(X, d)$ *is isomorphic to G.*

Proof. Let p be a bounded left-invariant compatible metric on G (if G is discrete, we may additionally assume that p is complete). It is an easy exercise (and a well-known fact) that all left translations on G form a closed subgroup of $\mathrm{Iso}(G, p)$. Consequently, by Theorem 3.2, G is isomorphic to $\mathrm{Iso}(Y, \varrho)$ for some metric ϱ which respects p. Note that if G is discrete, Y is homeomorphic to G, which proves (A). Further, if G is countable and nondiscrete, it is homeomorphic to the space of all rationals (for example, by Sierpiński's theorem [1920] that every countable metrizable topological space without isolated points is homeomorphic to $\mathbb{Q}$; see point (d) of Exercise 6.2.A on page 370 in [Engelking 1989]) and hence Y is homeomorphic to F, from which we deduce (B). Finally, (C) simply follows from the fact that, if G is totally disconnected, or zero-dimensional, or strongly zero-dimensional, then Y has the same property: see Theorems 1.3.6, 4.1.25, and 4.1.3 in [Engelking 1978]. □

Corollary 4.8(A) may be generalized to the context of so-called nonarchimedean complete topological groups. Recall that a topological group is *nonarchimedean* if and only if it has a base of neighborhoods of the neutral element consisting of open subgroups. Nonarchimedean Polish groups play an important role, for example, in model theory; see §1.5 in [Becker and Kechris 1996]. The equivalence between points (i) and (ii) of the following result is taken from this book (it was formulated there only for Polish groups, but the proof works in the general case):

Corollary 4.9. *Let G be a topological group and $\beta = w(G)$. Let D be a discrete topological space of cardinality β and S_β be the symmetric group of the set D (that is, S_β consists of all permutations of D). The following conditions are equivalent*:

(i) *G is nonarchimedean and complete.*

(ii) *G is isomorphic to a closed subgroup of S_β.*

(iii) *For arbitrary $\varepsilon > 0$, G is isomorphic to $\mathrm{Iso}(D, \varrho)$ for some metric ϱ on D such that $\delta_D \leqslant \varrho \leqslant (1+\varepsilon)\delta_D$, where δ_D is the discrete metric on D.*

Proof. First note that $S_\beta = \mathrm{Iso}(D, \delta)$. Thus, if (ii) is satisfied, Theorem 3.2 implies that there is a metric ϱ on $X = \widehat{D}_\beta$ which respects δ_D such that G is isomorphic to $\mathrm{Iso}(X, \varrho)$. Since ϱ respects δ_D, we readily have $\varrho \geqslant \delta_X$. Use Remark 3.3 to improve the metric ϱ and take care of the inequality $\varrho \leqslant (1+\varepsilon)\delta_X$. Finally, noticing that X is homeomorphic to D gives (iii).

Now assume (iii) is satisfied. Since ϱ is complete, so is G, by Proposition 4.3. Furthermore, the sets $U_A = \{u \in \mathrm{Iso}(D, \varrho) \mid u(a) = a \text{ for any } a \in A\}$, where A runs over all finite subsets of D, are open (because $\varrho \geqslant 1$) subgroups of $\mathrm{Iso}(D, \varrho)$ which form a base of neighborhoods of the identity map and thus G is nonarchimedean.

Finally, assume (i) is fulfilled. Let $\mathcal{B}$ be a base of neighborhoods of the neutral element e_G of G which consists of open subgroups and has cardinality β. For any $H \in \mathcal{B}$, the size of the collection $\{gH \mid g \in G\}$ does not exceed β and hence $\mathrm{card}(\mathcal{U}) = \beta = \mathrm{card}(D)$ for $\mathcal{U} = \{gH \mid g \in G, H \in \mathcal{B}\}$. Hence we may and do identify the set D with $\mathcal{U}$. For any $g \in G$ put $\pi_g : \mathcal{U} \ni U \mapsto gU \in \mathcal{U}$. Under the above identification, we readily have $\pi_g \in S_\beta$. What is more, the function $\pi : G \ni g \mapsto \pi_g \in S_\beta$ is easily seen to be a group homomorphism (possibly discontinuous). It suffices to check that π is an embedding (because then $\pi(G)$ is closed, thanks to the completeness of G). Since $\bigcap \mathcal{B} = \{e_G\}$, π is one-to-one. To complete the proof, observe that for any net $\{g_\sigma\}_{\sigma\in\Sigma} \subset G$ one has

$$\begin{aligned}\lim_{\sigma\in\Sigma} g_\sigma = e_G &\iff \forall g \in G : \lim_{\sigma\in\Sigma} g^{-1} g_\sigma g = e_G \\ &\iff \forall g \in G \ \ \forall H \in \mathcal{B} \ \ \exists \sigma_0 \in \Sigma \ \ \forall \sigma \geqslant \sigma_0 : \pi_{g_\sigma}(gH) = gH \\ &\iff \lim_{\sigma\in\Sigma} \pi_{g_\sigma} = \pi_{e_G}\end{aligned}$$

(recall that $\mathscr{U}$, as identified with D, has discrete topology and that for any $x \in G$ and $H \in \mathscr{B}$, $x \in H$ if and only if $xH = H$). □

Using Proposition 3.4, we may easily strengthen Theorem 1.4:

Proposition 4.10. *For any $\mathscr{G}_\delta$-complete group G there are a complete metric space (X, d) with $w(X) = w(G)$, a dense set $X' \subset X$, and an isomorphism $\Phi : \overline{G} \to \operatorname{Iso}(X, d)$ such that $\Phi(G) = \{u \in \operatorname{Iso}(X, d) \mid u(X') = X'\}(= \operatorname{Iso}(X', d))$.*

Proof. First use Theorem 1.4 to get the isomorphism between G and $\operatorname{Iso}(Y, \varrho)$ for some metric space (Y, ϱ) with $w(Y) = w(G)$ and then apply Proposition 3.4 to conclude the whole assertion (recall that the isometry group of a complete metric space is complete). □

Example 4.11. Let (X, d) be an arbitrary metric space and G be a subgroup of $\operatorname{Iso}(X, d)$. It follows from Theorem 1.4 and Proposition 4.1 that G is isomorphic to the isometry group of some metric space if and only if G is $\mathscr{G}_\delta$-closed in $\operatorname{Iso}(X, d)$. Let us briefly show that the $\mathscr{G}_\delta$-closure of G coincides with the set of all $u \in \operatorname{Iso}(X, d)$ such that for any separable subspace A of X there is $v \in G$ which agrees with u on A. Indeed, there is a countable set $D \subset A$ which is dense in A. Then the set $F(u, A) := \{v \in \operatorname{Iso}(X, d) : v|_A = u|_A\}$ coincides with $\{v \in \operatorname{Iso}(X, d) : v|_D = u|_D\}$. This implies that $F(u, A)$ is $\mathscr{G}_\delta$ in $\operatorname{Iso}(X, d)$. Consequently, if u belongs to the $\mathscr{G}_\delta$-closure of G then necessarily $G \cap F(u, A) \neq \varnothing$. Conversely, it may be easily shown that for every $\mathscr{G}_\delta$-set P containing u there is a countable set A such that $F(u, A) \subset P$ and hence to conclude that u belongs to the $\mathscr{G}_\delta$-closure of G it is sufficient that $F(u, A) \cap G \neq \varnothing$ for all countable A.

According to the above remark, Theorem 3.2 may now be generalized as follows: a subgroup G of $\operatorname{Iso}(X, d)$ is isomorphic to the isometry group of some metric space (of the same topological weight as X) if and only if G satisfies the following condition: Whenever $u \in \operatorname{Iso}(X, d)$ is such that for any separable subspace A of X there exists $v \in G$ which agrees with u on A then $u \in G$.

We end the section with the concept of $\mathscr{G}_\delta$-completions. Similarly, as in the case of Raĭkov completion, any topological group G has a unique $\mathscr{G}_\delta$-completion; that is, G may be embedded in a $\mathscr{G}_\delta$-complete group as a $\mathscr{G}_\delta$-dense subgroup in a unique way, as shown by:

Proposition 4.12. *Let G and K be $\mathscr{G}_\delta$-complete groups and H be a $\mathscr{G}_\delta$-dense subgroup of G:*

(a) *Every continuous homomorphism of H into K extends uniquely to a continuous homomorphism of G into K.*

(b) *If $f : H \to K$ is a group homomorphism as well as a topological embedding and $\widetilde{G}$ denotes the $\mathscr{G}_\delta$-closure of $f(H)$ in K then there is a unique (topological) isomorphism $F : G \to \widetilde{G}$ that extends f.*

Proof. Let $f : H \to K$ be a continuous homomorphism. Since H is dense in G, there is a unique continuous group homomorphism $F : G \to \overline{K}$ which extends f. It suffices to show that $F(G) \subset K$. But this follows from the fact that the preimage of a $\mathscr{G}_\delta$-closed set under a continuous function is $\mathscr{G}_\delta$-closed too. This proves (a). If in addition f is a topological embedding, it follows from the above argument that there is a continuous group homomorphism $\widetilde{F} : \widetilde{G} \to G$ which extends f^{-1}. We then readily see that both $\widetilde{F} \circ F$ and $F \circ \widetilde{F}$ are the identity maps and hence F is an isomorphism (and $\widetilde{F} = F^{-1}$), which shows (b). □

Definition 4.13. Let G be a topological group. The $\mathscr{G}_\delta$*-completion* of G is a $\mathscr{G}_\delta$-complete group which contains G as a $\mathscr{G}_\delta$-dense (topological) subgroup. It follows from Proposition 4.12 that the $\mathscr{G}_\delta$-completion is unique up to isomorphism fixing the points of G. It is also obvious that any group has the $\mathscr{G}_\delta$-completion.

5. Hilbert spaces as underlying topological spaces

Our first aim of this section is to prove Theorem 1.2 and Proposition 1.5(b). To this end, we recall a classical construction due to Arens and Eells [1956] (see also [Weaver 1999, Chapter 2]).

Definition 5.1. Let (X, d) be a nonempty complete metric space. For every $p \in X$ let $\chi_p : X \to \{0, 1\}$ be such that $\chi_p(x) = 1$ if $x = p$ and $\chi_p(x) = 0$ otherwise. A *molecule* of X is any function $m : X \to \mathbb{R}$ which is supported on a finite set and satisfies $\sum_{p\in X} m(p) = 0$. Denote by $\mathrm{AE}_0(X)$ the real vector space of all molecules of X, and for $m \in \mathrm{AE}_0(X)$ put

$$\|m\|_{AE} = \inf\left\{\sum_{j=1}^{n} |a_j| d(p_j, q_j) \,\middle|\, m = \sum_{j=1}^{n} a_j(\chi_{p_j} - \chi_{q_j})\right\}.$$

Then $\|\cdot\|_{AE}$ is a norm and the completion of $(\mathrm{AE}_0(X), \|\cdot\|_{AE})$ is called the *Arens–Eells* space of (X, d) and denoted by $(\mathrm{AE}(X), \|\cdot\|_{AE})$. Moreover, $w(\mathrm{AE}(X)) = w(X)$.

It is an easy observation that every isometry $u : X \to Y$ between complete metric spaces X and Y induces a unique linear isometry $\mathrm{AE}(u) : \mathrm{AE}(X) \to \mathrm{AE}(Y)$ such that $\mathrm{AE}(u)(\chi_p - \chi_q) = \chi_{u(p)} - \chi_{u(q)}$ for any $p, q \in X$. The following result is surely well-known, but probably nowhere explicitly stated. Therefore, for the reader's convenience, we give its short proof:

Lemma 5.2. *For every complete metric space* (X, d), *the function*

$$\Psi : \mathrm{Iso}(X, d) \ni u \mapsto \mathrm{AE}(u) \in \mathrm{Iso}(\mathrm{AE}(X), \|\cdot\|_{AE}) \tag{30}$$

is both a group homomorphism and a topological embedding.

Proof. Continuity and homomorphicity of Ψ is clear (note that $\mathrm{AE}_0(X)$ is dense in $\mathrm{AE}(X)$ and $\mathrm{AE}_0(X)$ is the linear span of the set $\{\chi_p - \chi_q \mid p, q \in X\}$). Here we focus only on showing that Ψ is an embedding. We may and do assume that $\operatorname{card}(X) > 1$. Suppose $\{u_\sigma\}_{\sigma\in\Sigma} \subset \operatorname{Iso}(X, d)$ is a net such that

$$\lim_{\sigma\in\Sigma} \Psi(u_\sigma) = \Psi(u) \tag{31}$$

for some $u \in \operatorname{Iso}(X, d)$. Let $x \in X$ be arbitrary. We only need to verify that $\lim_{\sigma\in\Sigma} u_\sigma(x) = u(x)$. Let $y \in X$ be different from x. We infer from (31) that

$$\chi_{u_\sigma(x)} - \chi_{u_\sigma(y)} \to \chi_{u(x)} - \chi_{u(y)} \quad (\sigma \in \Sigma). \tag{32}$$

For $\varepsilon \in (0, d(x, y))$ let $B_\varepsilon = \{z \in X \mid d(u(x), z) \geqslant \varepsilon\}(\neq \varnothing)$ and let $v_\varepsilon : X \to [0, \infty)$ denote the distance function to B_ε, that is,

$$v_\varepsilon(z) = \inf\{d(z, b) \mid b \in B_\varepsilon\}.$$

Observe that v_ε is Lipschitz. Since the dual Banach space to $\mathrm{AE}(X)$ is naturally isomorphic to the Banach space of all Lipschitz real-valued functions on X (see, for example, Chapter 2 of [Weaver 1999]), we deduce from (32) that $\lim_{\sigma\in\Sigma}[v_\varepsilon(u_\sigma(x)) - v_\varepsilon(u_\sigma(y))] = v_\varepsilon(u(x)) - v_\varepsilon(u(y))$. But $v_\varepsilon(u(x)) - v_\varepsilon(u(y)) = v_\varepsilon(u(x)) > 0$ (because $u(y) \in B_\varepsilon$, by the isometricity of u). So there is $\sigma_0 \in \Sigma$ such that $v_\varepsilon(u_\sigma(x)) > 0$ for any $\sigma \geqslant \sigma_0$. This means that for $\sigma \geqslant \sigma_0$, $u_\sigma(x) \notin B_\varepsilon$ and consequently $d(u_\sigma(x), u(x)) < \varepsilon$. □

The homomorphism appearing in (30) is *not* surjective, unless $\operatorname{card}(X) < 3$. There is however a fascinating result, discovered by Mayer–Wolf [1981], that characterizes *all* isometries of the space $\mathrm{AE}(X)$ under some additional conditions on the metric of X. Below we formulate only a special case of it, enough for our considerations.

Theorem 5.3. *Let d be a* ***bounded*** *complete metric on a set X. Let* $\mathrm{AE}(X)$ *denote the Arens–Eells spaces of* $(X, \sqrt{d})$. *Every linear isometry of* $\mathrm{AE}(X)$ *onto itself is of the form* $\pm\,\mathrm{AE}(u)$, *where* $u \in \operatorname{Iso}(X, d)(= \operatorname{Iso}(X, \sqrt{d}))$.

Proof. All parts of this proof come from [Mayer-Wolf 1981]. Alternatively, we give references to suitable results from [Weaver 1999]. By Proposition 2.4.5 of that reference, the metric space $(X, \sqrt{d})$ is so-called concave (for the definition, see the note preceding Lemma 2.4.4 on page 51 in [Weaver 1999]). Now, Theorem 2.7.2 in that reference implies that if Φ is a linear isometry of $\mathrm{AE}(X)$ then there is $r \in \mathbb{R} \setminus \{0\}$ and a bijection $u : X \to X$ such that $\sqrt{d(u(x), u(y))} = |r|\sqrt{d(x, y)}$ and $\Phi(\chi_x - \chi_y) = \frac{1}{r}(\chi_{u(x)} - \chi_{u(y)})$ for any $x, y \in X$. The former of these relations, combined with the boundedness of d, implies that $|r| = 1$ and $u \in \operatorname{Iso}(X, d)$. So $\Phi = \pm\,\mathrm{AE}(u)$, and we are done. □

We shall also need quite an intuitive result stated below. Although its proof is not immediate, we leave it to the reader as an exercise.

Lemma 5.4. *Let X be a two-dimensional real vector space, $\|\cdot\|$ be any norm on X, and let a and b be two vectors in X:*

(a) *If $\bar{B}_X(0, 2) \subset \bar{B}_X(b, 2) \cup \bar{B}_X(a, 1)$ then $b = 0$.*

(b) *If $\|a\| = \|b\| = 2$ and $\bar{B}_X(b, 1) \subset \bar{B}_X(0, 2) \cup \bar{B}_X(a, 1)$ then $a = b$.*

Proof of Theorem 1.2 and of Proposition 1.5(b). Because Theorem 1.2 is a special case of point (b) of the proposition, we focus only on the proof of the latter result. It follows from Corollary 4.7 that there is a complete metric space (Y, ϱ) such that $w(Y) = \beta$ and $\mathrm{Iso}(Y, \varrho)$ is isomorphic to G. Since $\mathrm{Iso}(Y, \varrho) = \mathrm{Iso}(Y, \varrho/(2+2\varrho))$ and the metric $\varrho/(2+2\varrho)$ is complete (and compatible), we may and do assume that $\varrho < \frac{1}{2}$. We also assume that $Y \cap [0, 1] = \varnothing$. Let $X = Y \cup [0, 1]$. We define a metric d on X by the rules:

- $d(s, t) = |s - t|$ for $s, t \in [0, 1]$.
- $d(x, y) = \varrho(x, y)$ for $x, y \in Y$.
- $d(x, t) = d(t, x) = 1 + t$ for $x \in Y$ and $t \in [0, 1]$.

We leave it as an exercise that d is indeed a metric, that d is complete, and $w(X) = \beta$. Notice that for any $a \in X$ and $t \in [0, 1] (\subset X)$:

- $a = 1 \iff \exists b, c \in X : d(a, b) = \frac{3}{4} \wedge d(a, c) = 2.$
- $a = t \iff d(a, 1) = 1 - t.$

These equivalences imply that for every $f \in \mathrm{Iso}(X, d)$ we have $f(t) = t$ for $t \in [0, 1]$ and $f|_Y \in \mathrm{Iso}(Y, \varrho)$. It is also easy to see that each isometry of (Y, ϱ) extends (uniquely) to an isometry of (X, d). Hence the function $\mathrm{Iso}(X, d) \ni f \mapsto f|_Y \in \mathrm{Iso}(Y, \varrho)$ is a (well-defined) isomorphism. Now let $(E, \|\cdot\|) = (\mathrm{AE}(X), \|\cdot\|_{AE})$ be the Arens–Eells space of $(X, \sqrt{d})$ and let $e = \chi_1 - \chi_0$. We see that $w(E) = \beta$ and E is infinite-dimensional, since X is infinite. What is more, it follows from Theorem 5.3 that every linear isometry of E which leaves the point e fixed is of the form $\mathrm{AE}(u)$ for some $u \in \mathrm{Iso}(X, \sqrt{d})$. Since $\mathrm{AE}(u)(e) = e$ for any $u \in \mathrm{Iso}(X, d)$ (because $u(0) = 0$ and $u(1) = 1$ for such u), noticing that $\mathrm{Iso}(X, d) = \mathrm{Iso}(X, \sqrt{d})$ and Lemma 5.2 finishes the proof. $\square$

Our next aim is to show Theorem 1.1(a) and Proposition 1.5(a). We shall need the next three results.

Theorem 5.5 [Mankiewicz 1972]. *Whenever X and Y are normed vector spaces, U and V are connected open subsets of, respectively, X and Y then every isometry of U onto V extends to a unique affine isometry of X onto Y.*

We recall that a function $\Phi : X \to Y$ between real vector spaces X and Y is *affine* if $\Phi - \Phi(0)$ is linear.

The following result is a consequence, for example, of [Bessaga and Pełczyński 1975, Theorem VI.6.2] and a famous theorem of Toruńczyk's [1981; 1985], which says that every Banach space is homeomorphic to a Hilbert space:

Theorem 5.6. *Every closed convex set in an infinite-dimensional Banach space whose interior is nonempty is homeomorphic to an infinite-dimensional Hilbert space.*

Our last tool is the next result, which in the separable case was proved by Mogilski [1979]. The argument there works also in the nonseparable case. This theorem in its full generality may also be briefly concluded from the results of [Toruńczyk 1981; 1985]. For the reader's convenience, we give a sketch of its proof.

Theorem 5.7. *Let X be a metrizable space. If X is the union of its two closed subsets A and B such that each of A, B, and $A \cap B$ is homeomorphic to an infinite-dimensional Hilbert space $\mathscr{H}$ then X itself is homeomorphic to $\mathscr{H}$.*

Proof. Put $C = A \cap B$.

First assume that C is a Z-set in both A and B (for the definition of a Z-set see Section 6). Then there exist homeomorphisms $h_A : A \to \mathscr{H} \times (-\infty, 0]$ and $h_B : B \to \mathscr{H} \times [0, \infty)$ which coincide on C and $h_A(C) = h_B(C) = \mathscr{H} \times \{0\}$ (this follows from the theorem on extending homeomorphisms between Z-sets in Hilbert manifolds, see [Anderson 1967; Anderson and McCharen 1970; Chapman 1971; Bessaga and Pełczyński 1975, Chapter V]. Now it suffices to define $h : X \to \mathscr{H} \times \mathbb{R}$ as the union of h_A and h_B to obtain the homeomorphism we searched for.

Now we consider a general case. Let $X' = (A \times [-1, 0]) \cup (B \times [0, 1]) \subset X \times [-1, 1]$. Observe that $A' = A \times [-1, 0]$, $B' = B \times [0, 1]$, and $C' = C \times \{0\}$ are homeomorphic to $\mathscr{H}$ (by the assumptions of the theorem) and C' is a Z-set in both A' and B'. Thus, we infer from the first part of the proof that X' is homeomorphic to $\mathscr{H}$. Finally, note that the function $X' \ni (x, t) \mapsto (x, 0) \in X \times \{0\}$ is a proper retraction. So Toruńczyk's result [1981] implies that X is a manifold modeled on $\mathscr{H}$. Since it is contractible, it is homeomorphic to $\mathscr{H}$. $\square$

Proof of Theorem 1.1(a) *and Proposition 1.5*(a). Again, observe that point (a) of the theorem under the question is a special case of point (a) of the proposition. Therefore we focus only on the latter result. Let E and e be as in point (b) of Proposition 1.5. Replacing, if needed, e by $2e/\|e\|$, we may assume that $\|e\| = 2$. Denote by $\mathscr{E}$ the group of all linear isometries which leave e fixed. Let $W = \bar{B}_E(0, 2) \cup \bar{B}_E(e, 1)$ be equipped with the metric p induced by the norm of E. Notice that if $V \in \mathscr{E}$ then $V(W) = W$ and $V|_W \in \mathrm{Iso}(W, p)$. Conversely, for each $g \in \mathrm{Iso}(W, p)$, $g = V|_W$ for some linear isometry $V \in \mathscr{E}$. Let us briefly justify this claim. Let $x = g(0)$ and $y = g(e)$. Then $W \subset \bar{B}_E(x, 2) \cup \bar{B}_E(y, 1)$ and consequently $\bar{B}_X(0, 2) \subset \bar{B}_X(x, 2) \cup \bar{B}_X(y, 1)$, where X is a two-dimensional linear subspace of E which contains x and y. We infer from Lemma 5.4(a) that $x = 0$. So $\|y\| = 2$ (since g is an isometry) and thus $\bar{B}_Y(e, 1) \subset \bar{B}_Y(0, 2) \cup \bar{B}_Y(y, 1)$, where Y is a two-dimensional linear subspace of E such that $e, y \in Y$. Now point (b) of Lemma 5.4 yields that $y = e$. We then have $g(B_E(0, 2) \cup B_E(e, 1)) = B_E(0, 2) \cup B_E(e, 1)$. So an application of Theorem 5.5 gives our assertion: There is a linear (since $g(0) = 0$) isometry V of E which extends g.

Having the above fact, we easily see that the function $\mathscr{E} \ni V \mapsto V|_W \in \mathrm{Iso}(W, p)$ is an isomorphism. Consequently, G is isomorphic to $\mathrm{Iso}(W, p)$. So, to finish the proof, it suffices to show that W is homeomorphic to $\mathscr{H}_\beta$. But this immediately follows from Theorems 5.6 and 5.7, since $w(E) = \beta$ and the sets $\bar{B}_E(0, 2)$, $\bar{B}_E(e, 1)$, and $\bar{B}_E(0, 2) \cap \bar{B}_E(1, 2)$ are closed, convex, and have nonempty interiors. □

Proof of Corollary 1.6. It suffices to apply Proposition 1.5 and Proposition 4.3(f). □

The arguments used in the proofs of both points of Proposition 1.5 also show the next result:

Corollary 5.8. *Let G be a $\mathscr{G}_\delta$-complete topological group of topological weight not exceeding $\beta \geqslant \aleph_0$. There are an infinite-dimensional normed vector space E of topological weight β, a contractible open set $U \subset E$, and a nonzero vector $e \in E$ such that the topological groups G, $\mathrm{Iso}(U, d)$, and $\mathrm{Iso}(E|e)$ are isomorphic, where d is the metric on U induced by the norm of E and $\mathrm{Iso}(E|e)$ is the group of all linear isometries of E which leave the point e fixed.*

Proof. Let (Y_0, ϱ_0) be a metric space such that $w(Y_0) = \beta$ and $\mathrm{Iso}(Y_0, \varrho_0)$ is isomorphic to G (see Theorems 1.4 and 3.2). Denote by (Y, ϱ) the completion of (Y_0, ϱ_0). Now let (X, d), $(\mathrm{AE}(X), \|\cdot\|_{AE})$, and e be as in the proof of Proposition 1.5(b). We know that the function

$$\mathrm{Iso}(X, d) \ni u \mapsto \mathrm{AE}(u) \in \mathrm{Iso}(\mathrm{AE}(X)|e)$$

is an isomorphism. Denote by E the linear span of the set $\{\chi_a - \chi_b \mid a, b \in Y_0 \cup [0, 1]\} (\subset \mathrm{AE}(X))$ (recall that $X = Y \sqcup [0, 1]$). Observe that if $u \in \mathrm{Iso}(X, d)$ is such that $\mathrm{AE}(u)(E) = E$ then $u(Y_0 \cup [0, 1]) = Y_0 \cup [0, 1]$ and consequently $u(Y_0) = Y_0$ (see the proof of point (b) of Proposition 1.5). This yields that the function

$$\mathrm{Iso}(Y_0, \varrho_0) \ni u \mapsto \mathrm{AE}(u)|_E \in \mathrm{Iso}(E|e)$$

is also an isomorphism. Now it suffices to put $U = B_E(0, \|e\|) \cup B_E(e, \frac{1}{2}\|e\|)$ and repeat the proof of Proposition 1.5(a) (involving Lemma 5.4 and Theorem 5.5) to get the whole assertion. (U is contractible as the union of two intersecting convex sets.) □

6. Isometry groups of completely metrizable metric spaces

Taking into account Corollary 1.6, the following question naturally arises: Given an infinite cardinal β, how do we characterize topological groups isomorphic to $\mathrm{Iso}(\mathcal{H}, d)$ for some compatible metric d on a Hilbert space $\mathcal{H}$ of Hilbert space dimension β? In this part we give a partial answer to this question. In fact, we will deduce our main result in this topic from the following general fact:

Proposition 6.1. *Let (S, p) be a bounded complete metric space, β an infinite cardinal not less than $w(S)$, and $\mathcal{H}$ a Hilbert space of Hilbert space dimension β. Let S' be a dense subset of S and G a closed subgroup of* $\mathrm{Iso}(S', p)$. *There exist a compatible complete metric λ on $\mathcal{H}$, a set $\mathcal{H}' \subset \mathcal{H}$, and an isomorphism $\Psi : \overline{G} \to \mathrm{Iso}(\mathcal{H}, \lambda)$ such that $\Psi(G) = \{u \in \mathrm{Iso}(\mathcal{H}, \lambda) \mid u(\mathcal{H}') = \mathcal{H}'\}$, $(\mathcal{H} \setminus \mathcal{H}', \lambda)$ is isometric to $(S \setminus S', \sqrt{p})$ and the closure of $\mathcal{H} \setminus \mathcal{H}'$ is a Z-set in $\mathcal{H}$. In particular, if S' is completely metrizable then $\mathcal{H}'$ is homeomorphic to $\mathcal{H}$.*

Recall that a closed set K in a metric space X is a *Z-set* if and only if every map of the Hilbert cube Q into X may uniformly be approximated by maps of Q into $X \setminus K$.

Proof. By Proposition 3.4, there is a bounded complete metric space (Y, ϱ), a dense set $Y' \subset Y$, and an isomorphism $F_1 : \overline{G} \to \mathrm{Iso}(Y, \varrho)$ such that $w(Y) = \beta$, $(Y \setminus Y', \varrho)$ is isometric to $(S \setminus S', p)$, and $F_1(G) = \{u \in \mathrm{Iso}(Y, \varrho) \mid u(Y') = Y'\}$. Now we shall mimic the proof of Proposition 1.5.

Replacing, if applicable, p and ϱ by $t \cdot p$ and $t \cdot \varrho$ with small enough $t > 0$ (and the final metric λ obtained from this proof by $\lambda/\sqrt{t}$), we may assume that $\varrho < \frac{1}{2}$. Now let $(X, d) \supset (Y, \varrho)$ be as in the proof of Theorem 1.2. Further, let $(E, \|\cdot\|) = (\mathrm{AE}(X), \|\cdot\|_{AE})$ be the Arens–Eells space of $(X, \sqrt{d})$ and $e = \chi_1 - \chi_0 \in E$. We denote by λ the metric on $W = \bar{B}_E(0, 1) \cup \bar{B}_E(e, \frac{1}{2})$ induced by the norm of E. The arguments used in the proofs of Theorem 1.2 and Proposition 1.5 show:

(I1) The function $F_2 : \mathrm{Iso}(X, d) \ni u \mapsto u|_Y \in \mathrm{Iso}(Y, \varrho)$ is an isomorphism and there is a dense set $X' \subset X$ such that $X \setminus X' = Y \setminus Y'$ and $F_2^{-1}(F_1(G))$ consists precisely of all $u \in \mathrm{Iso}(X, d)$ with $u(X') = X'$.

(I2) The function $F_3 : \mathrm{Iso}(X, d) \ni u \mapsto \mathrm{AE}(u)|_W \in \mathrm{Iso}(W, \lambda)$ is an isomorphism and $u(0) = 0$ for any $u \in \mathrm{Iso}(X, d)$.

(I3) W is homeomorphic to $\mathcal{H}$.

Point (I3) asserts that we may identify $\mathcal{H}$ with W. Put $\Psi = F_3 \circ F_2^{-1} \circ F_1 : \overline{G} \to \mathrm{Iso}(W, \lambda)$ and $W' = W \setminus \{\chi_x - \chi_0 \mid x \in X \setminus X'\}$ (recall that $\|\chi_y - \chi_0\|_{AE} = \sqrt{d(y, 0)} = 1$ for every $y \in Y \supset X \setminus X'$) and note that Ψ is an isomorphism. What is more, we claim that $\Psi(G)$ consists of all $v \in \mathrm{Iso}(W, \lambda)$ for which $v(W') = W'$. Indeed, it follows from (I2) that each $v \in \mathrm{Iso}(W, \lambda)$ has the form $v = \mathrm{AE}(u)|_W$ for some

$u \in \mathrm{Iso}(X, d)$. So, taking into account (I1), we only need to check that $u(X') = X'$ if and only if $(\mathrm{AE}(u))(W') = W'$. But this follows from the fact that $u(0) = 0$.

Further, observe that the function $(X \setminus X', \sqrt{d}) \ni x \mapsto \chi_x - \chi_0 \in (W \setminus W', \lambda)$ is an isometry, which implies that the latter metric space is isometric to $(S \setminus S', \sqrt{p})$. To show that the closure D of $W \setminus W'$ is a Z-set in W, note that $D \subset C := \{\chi_y - \chi_0 \mid y \in Y\}$ and the maps $W \ni w \mapsto (1 - 1/n)w + e/n \in W$ converge uniformly to the identity map with respect to λ and their images are disjoint from C.

So, to complete the proof, we only need to check that if S' is completely metrizable then W' is homeomorphic to $\mathcal{H}$. But this is an immediate consequence of the fact that $W \setminus W'$ is a set isometric to $(S \setminus S', \sqrt{p})$ and contained in a Z-set in W, and a known fact on σ-Z-sets in Hilbert spaces: S', being completely metrizable, is a $\mathcal{G}_\delta$-set in S; hence, $S \setminus S'$ is $\mathcal{F}_\sigma$ in S. Consequently, $W \setminus W'$ is a countable union of sets complete in the metric λ thus an $\mathcal{F}_\sigma$-set in W. But the closure of $W \setminus W'$ is a Z-set in W, so $W \setminus W'$ is a σ-Z-set, that is, it is a countable union of Z-sets. Now the assertion follows from the well-known result that the complement of a σ-Z-set in a Hilbert space is homeomorphic to the whole space, which simply follows from Toruńczyk's characterization [1981; 1985] of Hilbert manifolds. (For the separable case one may also consult Theorem 6.3 in [Bessaga and Pełczyński 1975, Chapter V].) $\square$

As a conclusion, we obtain:

Theorem 6.2. *Let $\mathcal{H}$ be a Hilbert space of Hilbert space dimension $\beta \geqslant \aleph_0$ and*

$$\mathcal{G} = \{\mathrm{Iso}(\mathcal{H}, \varrho) \mid \varrho \text{ is a compatible metric on } \mathcal{H}\}.$$

Then, up to isomorphism, $\mathcal{G}$ consists precisely of all topological groups G which are isomorphic to closed subgroups of $\mathrm{Iso}(X, d)$ *for some completely metrizable spaces (X, d) with $w(X) \leqslant \beta$.*

Proof. If G is a closed subgroup of $\mathrm{Iso}(X, d)$ for a completely metrizable space (X, d) with $w(X) \leqslant \beta$, we may assume that d is bounded (replacing, if necessary, d by $d/(1 + d)$). Then Proposition 6.1, applied for $(S, p) =$ the completion of (X, d) and $S' = X$, yields a complete compatible metric λ on $\mathcal{H}$ and a dense set $\mathcal{H}' \subset \mathcal{H}$ homeomorphic to $\mathcal{H}$ such that G is isomorphic to $G' := \{u \in \mathrm{Iso}(\mathcal{H}, \lambda) \mid u(\mathcal{H}') = \mathcal{H}'\}$. But G' is naturally isomorphic to $\mathrm{Iso}(\mathcal{H}', \lambda)$ and we are done. $\square$

For an infinite cardinal number α, let us denote by $\mathsf{IGH}(\alpha)$ the class of all topological groups which are isomorphic to $\mathrm{Iso}(\mathcal{H}, \varrho)$ for some compatible metric ϱ on a Hilbert space $\mathcal{H}$ of Hilbert space dimension α ("IGH" is the abbreviation of "isometry group of a Hilbert space"). Additionally, let IGH stand for the union of all classes $\mathsf{IGH}(\alpha)$. Theorem 6.2 implies that IGH is a *variety*; that is, closed subgroups as well as topological products of members of IGH belong to IGH as well.

We are mainly interested in the class $\mathsf{IGH}(\aleph_0)$. It is clear that all groups belonging to this class are second-countable. In the sequel we shall see that the axiom of second countability is insufficient for a topological group to belong to $\mathsf{IGH}(\aleph_0)$ (see Proposition 6.5 below).

As a simple consequence of Theorem 6.2 we obtain:

Corollary 6.3. *If G is a second-countable, σ-compact topological group then $G \in \mathsf{IGH}(\aleph_0)$.*

Proof. Let p be a compatible left-invariant metric on G and let $(Y, \varrho) \supset (G, p)$ denote the completion of (G, p). If the interior of G in Y is nonempty then G is locally completely metrizable and thus G is Polish. In that case the assertion follows from Theorem 1.1. On the other hand, if G is a boundary set in Y then $Y \setminus G$ is dense in Y and therefore $\mathrm{Iso}(Y \setminus G, \varrho)$ is isomorphic to $\{u \in \mathrm{Iso}(Y, \varrho) \mid u(Y \setminus G) = Y \setminus G\}$ and the latter group coincides with $\{u \in \mathrm{Iso}(Y, \varrho) \mid u(G) = G\}$, which is isomorphic to $\mathrm{Iso}(G, \varrho) = \mathrm{Iso}(G, p)$. Since G is σ-compact, $Y \setminus G$ is completely metrizable. Finally, since p is left-invariant, all left translations of G form a closed subgroup of $\mathrm{Iso}(G, p)$ isomorphic to G. So, to sum up, G is isomorphic to a closed subgroup of the isometry group of the Polish space $(Y \setminus G, \varrho)$. Now it suffices to apply Theorem 6.2. $\square$

To formulate our next result, we remind the reader that a separable metrizable space X is said to be *coanalytic* if and only if X is homeomorphic to a space of the form $Y \setminus Z$, where Y is a Polish space and $Z \subset Y$ is *analytic*, that is, Z is the continuous image of a Polish metric space. We also recall that continuous images of Borel subsets of Polish spaces are analytic.

Proposition 6.4. *Each member of $\mathsf{IGH}(\aleph_0)$ is coanalytic as a topological space.*

Proof. Let (X, ϱ) be a Polish metric space and $G = \mathrm{Iso}(X, \varrho)$. Denote by (Y, d) the completion of (X, ϱ). Since X is completely metrizable, it is a $\mathscr{G}_\delta$-set in Y. Observe that G is naturally isomorphic to the subgroup $\{u \in \mathrm{Iso}(Y, d) \mid u(X) = X\}$ of $\mathrm{Iso}(Y, d)$. Since $\mathrm{Iso}(Y, d)$ is a Polish group, it suffices to show that $A := \{u \in \mathrm{Iso}(Y, d) \mid u(X) \neq X\}$ is the continuous image of a Borel subset of a Polish metric space. To this end, notice that the set $W := \{(u, x) \in \mathrm{Iso}(Y, d) \times X \mid u(x) \in Y \setminus X\}$ is Borel in the Polish space $\mathrm{Iso}(Y, d) \times X$ and $\pi(W) = A$, where $\pi : \mathrm{Iso}(Y, d) \times X \to \mathrm{Iso}(Y, d)$ is the natural projection. $\square$

Proposition 6.5. *There exists a topological subgroup of the additive group of reals which does not belong to $\mathsf{IGH}(\aleph_0)$.*

Proof. We consider $\mathbb{R}$ as a vector space over the field $\mathbb{Q}$ of rationals. There exists a vector subspace G of $\mathbb{R}$ such that $\mathbb{Q} \cap G = \{0\}$ and $G + \mathbb{Q} = \mathbb{R}$. We claim that $G \notin \mathsf{IGH}(\aleph_0)$. To show that, it is enough to prove that G is not coanalytic (thanks to Proposition 6.4). Since analytic spaces are absolutely measurable (see, for example, Theorem A.13 in [Takesaki 1979, Appendix]), it suffices to show that G is not Lebesgue measurable. But this follows from the following two observations:

- $\mathbb{R} = \bigcup_{q\in\mathbb{Q}}(q + G)$ and hence the outer measure of G is positive.
- $G - G$ (which equals G) has empty interior and thus its inner measure is 0; this follows from [Halmos 1950, Chapter XII, §61, Theorem A, p. 266]. $\square$

Propositions 6.4 and 6.5 make the issue of characterizing members of IGH complicated. The following problems seem to be most interesting:

Problem 6.6. If $G \in \mathsf{IGH}$, is it true that $G \in \mathsf{IGH}(w(G))$?

Problem 6.7. Does the class IGH contain all $\mathscr{G}_\delta$-complete topological groups?

Problem 6.8. Characterize members of $\mathsf{IGH}(\aleph_0)$.

7. Compact and locally compact Polish groups

This section is devoted to the proofs of points (b) and (c) of Theorem 1.1. Our main tool will be the following result, very recently shown by us in [Niemiec 2012]:

Theorem 7.1. *Let G be a locally compact Polish group, X be a locally compact Polish space, and let $G \times X \ni (g, x) \mapsto g.x \in X$ be a continuous proper action of G on X. Assume there is a point $\omega \in X$ such that the set $G.\omega = \{g.\omega \mid g \in G\}$ is nonopen and G acts freely at ω (that is, $g.\omega = \omega$ implies $g =$ the neutral element of G). Then there exists a proper compatible metric d on X such that* $\operatorname{Iso}(X, d)$ *consists precisely of all maps of the form $x \mapsto g.x$ ($g \in G$). In particular, the topological groups* $\operatorname{Iso}(X, d)$ *and G are isomorphic.*

We recall that (under the above notation) the action is *proper* if for every compact set $K \subset X$ the set $\{g \in G \mid g.K \cap K \neq \varnothing\}$ is compact as well (where $g.K = \{g.x \mid x \in K\}$).

Our next tool is the following classical result due to Keller [1931] (see also [Bessaga and Pełczyński 1975, Theorem III.3.1]):

Theorem 7.2. *Every infinite-dimensional compact convex subset of a Fréchet space is homeomorphic to the Hilbert cube.*

We recall that a Fréchet space is a completely metrizable locally convex topological vector space.

We call a function $u : (X, d) \to \mathbb{R}$ (where (X, d) is a metric space) *nonexpansive* if and only if $|u(x) - u(y)| \leqslant d(x, y)$ for all $x, y \in X$. The function u is a *Katětov map* if and only if u is nonexpansive and additionally $d(x, y) \leqslant u(x) + u(y)$ for any $x, y \in X$. Katětov maps correspond to one-point extensions of metric spaces.

Proof of Theorem 1.1(b). Let G be a compact Polish group. Take a left-invariant metric $\varrho \leqslant 1$ on G and equip the space $X = G \times [0, 1]$ with the metric d, where

$d((x,s),(y,t)) = \max(\varrho(x,y), |t-s|)$. For $g \in G$ denote by ψ_g the function $X \ni (x,t) \mapsto (g^{-1}x, t) \in X$. Notice that $\psi_g \in \mathrm{Iso}(G, X)$ for any $g \in G$. Let Δ be the space of all nonexpansive maps of (X, d) into $[0, 1]$ endowed with the supremum metric. Observe that Δ is a convex set in the Banach space of all real-valued maps on X. What is more, Δ is infinite-dimensional since X is infinite, and Δ is compact by the Ascoli type theorem. So we infer from Theorem 7.2 that Δ is homeomorphic to Q. Further, $\Phi_g(u) := u \circ \psi_g \in \Delta$ for any $g \in G$ and $u \in \Delta$ (because ψ_g is isometric). It is also easily seen that the function $G \times \Delta \ni (g,u) \mapsto \Phi_g(u) \in \Delta$ is a (proper — since both G and Δ are compact) continuous action of G on Δ. Finally, the function $\omega \colon X \ni (x,t) \mapsto d((x,t),(e,1)) \in \mathbb{R}$ belongs to Δ (since $d \leqslant 1$), where e is the neutral element of G. Observe that the set $K := \{\omega \circ \psi_g \mid g \in G\}$ has empty interior in Δ, since $1/n + (1 - 1/n)\omega \circ \psi_g \in \Delta \setminus K$ for any $n \geqslant 1$. Now we apply Theorem 7.1. □

Our last aim is to prove Theorem 1.1(c). To this end, we need more information on Hilbert cube manifolds.

One of the deepest results in infinite-dimensional topology is Anderson's theorem [1967] on extending homeomorphisms between Z-sets. Below we formulate it only in the Hilbert cube settings, it holds however in a much more general context. (For the discussion on this topic consult [Bessaga and Pełczyński 1975, Chapter V]; see also [Anderson and McCharen 1970; Chapman 1971]).

Theorem 7.3. *Every homeomorphism between two Z-sets in the Hilbert cube Q is extendable to a homeomorphism of Q onto itself.*

The result stated below is a kind of folklore in Hilbert cube manifolds theory. We present its short proof because we could not find it in the literature.

Theorem 7.4. *The spaces $Q \times [0, \infty)$ and $Q \setminus \{\text{point}\}$ are homeomorphic.*

Proof. Since $Q \setminus \{\text{point}\}$ is a Hilbert cube manifold, it follows from Schori's theorem [1971] (see also [Chapman 1976]; compare with [Bessaga and Pełczyński 1975, Theorem IX.4.1]) that $(Q \setminus \{\text{point}\}) \times Q$ is homeomorphic to $Q \setminus \{\text{point}\}$. Now the assertion follows from Theorem 7.3 since $(Q \times [0,1]) \setminus (Q \times [0,1)) = Q \times \{1\}$ is a Z-set in $Q \times [0,1]$ homeomorphic to the Z-set (in $Q \times Q$) $(Q \times Q) \setminus [(Q \setminus \{\text{point}\}) \times Q]$. □

Lemma 7.5. *Let (X, d) be a nonempty separable metric space and let $E(X)$ be the set of all Katětov maps on (X, d) equipped with the pointwise convergence topology*:

(i) *For any $a \in X$ and $r > 0$ the set $\{f \in E(X) \mid f(a) \leqslant r\}$ is compact (in $E(X)$).*

(ii) *$E(X) \times Q$ is homeomorphic to $Q \setminus \{\text{point}\}$.*

Proof. Point (i) follows from the Ascoli type theorem, since $E(X)$ consists of nonexpansive maps and for any $f \in E(X)$ and $x \in X$, $f(x) \in [0, d(x,a) + f(a)]$.

We turn to (ii). First of all, $E(X)$ is metrizable, because of the separability of X and the nonexpansivity of members of $E(X)$. Further, thanks to Theorem 7.4, it suffices to show that $E(X) \times Q$ is homeomorphic to $Q \times [0, \infty)$. Fix $a \in X$ and let $\omega \in E(X)$ be given by $\omega(x) = d(a, x)$. For each $n \geqslant 1$ let $K_n = \{f \in E(X) \mid f(a) \in [n-1, n]\}$ and $Z_{n-1} = \{f \in E(X) \mid f(a) = n-1\}$. We infer from (i) that K_n and Z_{n-1} are compact. It is also easily seen that both are convex nonempty sets ($\omega + n - 1 \in Z_{n-1} \subset K_n$). Since $K_n \times Q$ and $Z_{n-1} \times Q$ are affinely homeomorphic to convex subsets of Fréchet spaces, Theorem 7.2 yields that both these sets are homeomorphic to Q. Let $h_{n-1} : Z_{n-1} \times Q \to Q \times \{n-1\}$ be any homeomorphism. We claim that $Z_{n-1} \cup Z_n$ is a Z-set in K_n. This easily follows from the fact that the maps $K_n \ni f \mapsto (1 - 1/k)f + 1/k(\omega + n - \frac{1}{2}) \in K_n$ send K_n into $K_n \setminus (Z_{n-1} \cup Z_n)$ and converge uniformly (as $k \to \infty$) to the identity map of K_n. Since $Q \times \{n-1, n\}$ is a Z-set in $Q \times [n-1, n]$, Theorem 7.3 provides us the existence of a homeomorphism $H_n : K_n \times Q \to Q \times [n-1, n]$ which extends both h_{n-1} and h_n. We claim that the union $H : E(X) \times Q \to Q \times [0, \infty)$ of all H_n $(n \geqslant 1)$ is the homeomorphism we are searching for. It is clear that H is a well-defined bijection. Finally, notice that the interiors (in $E(X)$) of the sets $\bigcup_{j=1}^{n} K_j$ $(n \geqslant 1)$ cover X and hence H is indeed a homeomorphism. □

Proof of Theorem 1.1(c). Let G be a locally compact Polish group. By a theorem of Struble [1974] (see also [Abels et al. 2011]), there exists a proper left-invariant compatible metric d on G. Let $E(G)$ be the space of all Katětov maps on (G, d) endowed with the pointwise convergence topology. By Lemma 7.5, $L := E(G) \times Q$ is homeomorphic to $Q \setminus \{\text{point}\}$. So it suffices to show that there is a proper compatible metric ϱ on L such that $\mathrm{Iso}(L, \varrho)$ is isomorphic to G. For any $g \in G$ and $(f, q) \in L$ let $g.(f, q) = (f_g, q) \in L$, where $f_g(x) = f(g^{-1}x)$ (since d is left-invariant, $f_g \in E(G)$ for each $f \in E(G)$). As in the proof of point (b) of the theorem, we see that the function $G \times L \ni (g, x) \mapsto g.x \in L$ is a continuous action of G on L. It is also clear that each G-orbit (that is, each of the sets $G.x$ with $x \in L$) has empty interior. Similarly, as in point (b), we show that there is $\omega \in L$ such that G acts freely at ω (for example, $\omega = (u, q)$ with arbitrary $q \in Q$ and $u(x) = d(x, e)$, where e is the neutral element of G). So, by virtue of Theorem 7.1, it remains to check that the action is proper. To this end, take any compact set W in L. Then there is $r > 0$ such that $W \subset \{f \in E(G) \mid f(e) \leqslant r\} \times Q$. Note that the set $\{g \in G \mid g.W \cap W \neq \varnothing\}$ is closed and contained in $D \times Q$, where

$$D = \{g \in G \mid \exists f \in E(G) \text{ s.t. } f(e) \leqslant r \wedge f(g^{-1}) \leqslant r\}$$

and therefore it is enough to show that D has compact closure in G. But if $g \in D$, and $f \in E(G)$ is such that $f(e) \leqslant r$ and $f(g^{-1}) \leqslant r$ then $d(g, e) = d(e, g^{-1}) \leqslant f(e) + f(g^{-1}) \leqslant 2r$. This yields that $D \subset \bar{B}_G(e, 2r)$ and noting that d is proper finishes the proof. □

Remark 7.6. Van Dantzig and van der Waerden [1928] proved that the isometry group of a connected locally compact metric space (X, d) (possibly with nonproper or incomplete metric) is locally compact and acts properly on X. It follows from [Niemiec 2012] that there exists then a proper compatible metric ϱ on X such that $\mathrm{Iso}(X, d) = \mathrm{Iso}(X, \varrho)$. In particular,

$$\{\mathrm{Iso}(Q \setminus \{\text{point}\}, d) \mid d \text{ is a compatible metric}\} \\ = \{\mathrm{Iso}(Q \setminus \{\text{point}\}, \varrho) \mid \varrho \text{ is a proper compatible metric}\}$$

and hence if we omit the word *proper* in Theorem 1.1(c), we will obtain an equivalent statement.

As we mentioned in the introductory part, each (locally) compact finite-dimensional Polish group is isomorphic to the isometry group of a (proper locally) compact finite-dimensional metric space. Taking this, and Corollary 4.8, into account, the following question may be interesting:

Problem 7.7. Is every finite-dimensional metrizable (resp. finite-dimensional Polish) group isomorphic to the isometry group of a finite-dimensional (resp. finite-dimensional separable complete) metric space?

References

[Abels et al. 2011] H. Abels, A. Manoussos, and G. Noskov, "Proper actions and proper invariant metrics", *J. Lond. Math. Soc.* (2) **83**:3 (2011), 619–636. MR 2012e:54050 Zbl 1235.54017

[Anderson 1967] R. D. Anderson, "On topological infinite deficiency", *Michigan Math. J.* **14** (1967), 365–383. MR 35 #4893 Zbl 0148.37202

[Anderson and McCharen 1970] R. D. Anderson and J. D. McCharen, "On extending homeomorphisms to Fréchet manifolds", *Proc. Amer. Math. Soc.* **25** (1970), 283–289. MR 41 #2711 Zbl 0203.25805

[Arens and Eells 1956] R. F. Arens and J. Eells, Jr., "On embedding uniform and topological spaces", *Pacific J. Math.* **6** (1956), 397–403. MR 18,406e Zbl 0073.39601

[Arhangel'skii 2002] A. V. Arhangel'skii, "The Hewitt–Nachbin completion in topological algebra: Some effects of homogeneity", *Appl. Categ. Structures* **10**:3 (2002), 267–278. MR 2003f:54053 Zbl 1029.54033

[Arhangel'skii and Tkachenko 2008] A. Arhangel'skii and M. Tkachenko, *Topological groups and related structures*, Atlantis Studies in Mathematics **1**, Atlantis, Paris, 2008. MR 2010i:22001 Zbl 05293009

[Becker and Kechris 1996] H. Becker and A. S. Kechris, *The descriptive set theory of Polish group actions*, London Mathematical Society Lecture Note Series **232**, Cambridge University Press, 1996. MR 98d:54068 Zbl 0949.54052

[Berberian 1974] S. K. Berberian, *Lectures in functional analysis and operator theory*, Graduate Texts in Mathematics **15**, Springer, New York, 1974. MR 54 #5775 Zbl 0296.46002

[Bessaga and Pełczyński 1975] C. Bessaga and A. Pełczyński, *Lectures on Hilbert cube manifolds*, Mathematical Monographs **58**, Polish Scientific Publishers PWN, Warsaw, 1975. MR 57 #17657

[Chapman 1971] T. A. Chapman, "Deficiency in infinite-dimensional manifolds", *General Topology and Appl.* **1** (1971), 263–272. MR 48 #1259 Zbl 0228.57008

[Chapman 1976] T. A. Chapman, *Lectures on Hilbert cube manifolds*, C.B.M.S. Regional Conference Series in Math. **28**, American Mathematical Society, Providence, R. I., 1976. MR 54 #11336 Zbl 0347.57005

[Dantzig and Waerden 1928] D. v. Dantzig and B. L. v. d. Waerden, "Über metrisch homogene Räume", *Abh. Math. Sem. Hamburg* **6**:1 (1928), 367–376. JFM 54.0603.04

[Dieudonné 1939] J. Dieudonné, "Sur les espaces uniformes complets", *Ann. École Norm.* **56** (1939), 277–291. MR 1,220g Zbl 0024.36301

[Engelking 1978] R. Engelking, *Dimension theory*, North-Holland Mathematical Library **19**, North-Holland, Amsterdam, 1978. MR 58 #2753b Zbl 0401.54029

[Engelking 1989] R. Engelking, *General topology*, 2nd ed., Sigma Series in Pure Mathematics **6**, Heldermann Verlag, Berlin, 1989. MR 91c:54001 Zbl 0684.54001

[Gao and Kechris 2003] S. Gao and A. S. Kechris, *On the classification of Polish metric spaces up to isometry*, Mem. Amer. Math. Soc. **766**, American Mathematical Society, 2003. MR 2004b:03067 Zbl 1012.54038

[Halmos 1950] P. R. Halmos, *Measure theory*, Van Nostrand Company, New York, 1950. MR 11,504d Zbl 0040.16802

[Keller 1931] O.-H. Keller, "Die Homoiomorphie der kompakten konvexen Mengen im Hilbertschen Raum", *Math. Ann.* **105**:1 (1931), 748–758. MR 1512740 Zbl 0003.22401

[Malicki and Solecki 2009] M. Malicki and S. Solecki, "Isometry groups of separable metric spaces", *Math. Proc. Cambridge Philos. Soc.* **146**:1 (2009), 67–81. MR 2009j:54057 Zbl 1170.54013

[Mankiewicz 1972] P. Mankiewicz, "On extension of isometries in normed linear spaces", *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.* **20** (1972), 367–371. MR 47 #776 Zbl 0234.46019

[Mayer-Wolf 1981] E. Mayer-Wolf, "Isometries between Banach spaces of Lipschitz functions", *Israel J. Math.* **38**:1-2 (1981), 58–74. MR 82e:46036 Zbl 0459.46017

[Melleray 2008] J. Melleray, "Compact metrizable groups are isometry groups of compact metric spaces", *Proc. Amer. Math. Soc.* **136**:4 (2008), 1451–1455. MR 2008m:54048 Zbl 1134.54013

[Mogilski 1979] J. Mogilski, "CE-decomposition of l_2-manifolds", *Bull. Acad. Polon. Sci. Sér. Sci. Math.* **27**:3-4 (1979), 309–314. MR 81e:57018 Zbl 0429.57005

[Niemiec 2012] P. Niemiec, "Isometry groups of proper metric spaces", preprint, 2012. To appear in *Trans. Amer. Math. Soc.* arXiv 1201.5675

[Roelcke and Dierolf 1981] W. Roelcke and S. Dierolf, *Uniform structures on topological groups and their quotients*, McGraw-Hill, New York, 1981. MR 83j:22001 Zbl 0489.22001

[Schori 1971] R. Schori, "Topological stability for infinite-dimensional manifolds", *Compositio Math.* **23** (1971), 87–100. MR 44 #4789 Zbl 0219.57003

[Sierpiński 1920] W. Sierpiński, "Sur une propriété topologique des ensembles dénombrables denses en soi", *Fund. Math.* **1**:1 (1920), 11–16. JFM 47.0175.03

[Struble 1974] R. A. Struble, "Metrics in locally compact groups", *Compositio Math.* **28** (1974), 217–222. MR 50 #535 Zbl 0288.22010

[Takesaki 1979] M. Takesaki, *Theory of operator algebras, I*, Springer, New York, 1979. MR 81e:46038 Zbl 0436.46043

[Toruńczyk 1981] H. Toruńczyk, "Characterizing Hilbert space topology", *Fund. Math.* **111**:3 (1981), 247–262. MR 82i:57016 Zbl 0468.57015

[Toruńczyk 1985] H. Toruńczyk, "A correction of two papers concerning Hilbert manifolds", *Fund. Math.* **125**:1 (1985), 89–93. MR 87m:57017 Zbl 0199.43505

[Uspenskij 2008] V. V. Uspenskij, "On subgroups of minimal topological groups", *Topology Appl.* **155**:14 (2008), 1580–1606. MR 2009d:22003 Zbl 1166.22002

[Weaver 1999] N. Weaver, *Lipschitz algebras*, World Scientific, River Edge, NJ, 1999. MR 2002g:46002 Zbl 0936.46002

Received July 31, 2012. Revised February 16, 2013.

PIOTR NIEMIEC
INSTYTUT MATEMATYKI, WYDZIAŁ MATEMATYKI I INFORMATYKI
UNIWERSYTET JAGIELLOŃSKI
UL. ŁOJASIEWICZA 6
30-348 KRAKÓW
POLAND

piotr.niemiec@uj.edu.pl
www2.im.uj.edu.pl/PiotrNiemiec/publ.html

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.117

SINGULARITIES AND LIOUVILLE THEOREMS FOR SOME SPECIAL CONFORMAL HESSIAN EQUATIONS

QIANZHONG OU

We develop some new techniques to get an integral estimate for some special conformal Hessian equations, and hence the classification of their singularities. This complete results of González. By this method we were able to deduce the Liouville theorem for these special conformal Hessian equations, which were understood by Yanyan Li via the method of moving planes.

1. Introduction

Consider the conformal k-Hessian equation

$$\sigma_k(A^g) = u^\alpha \quad \text{in } \Omega, \tag{1-1}$$

where Ω is the whole space $\mathbb{R}^n$ or the punctured unit ball $B \setminus \{0\} \subset \mathbb{R}^n$ and $g = u^{-2}\,dx^2$, $u > 0$, is a locally conformally flat metric. The matrix A^g is given by $A^g = g^{-1}\widetilde{A}^g$, where $\widetilde{A}^g$ is the $(0, 2)$ Schouten tensor

$$\widetilde{A}^g_{ij} = \frac{1}{n-2}\Big(\mathrm{Ric}_{ij} - \frac{R}{2(n-1)} g_{ij}\Big),$$

where Ric and R denote the Ricci tensor and the scalar curvature of g, respectively. In this metric, the $(1, 1)$ Schouten tensor becomes

$$A^g = u(D^2u) - \tfrac{1}{2}|Du|^2 I. \tag{1-2}$$

These σ_k are k-Hessians of A^g. More precisely, they are defined as the k-th elementary symmetric polynomial functions of the eigenvalues $\lambda_1, \dots, \lambda_n$ of the symmetric matrix A^g:

$$\sigma_k(A^g) := \sum_{1\le i_1<\cdots<i_k\le n} \lambda_{i_1}\cdots\lambda_{i_k}.$$

The author's research was supported by NSFC grant 11061013 and the "Program for Excellent Talents in Guangxi Higher Education Institutions".
MSC2010: 35J60.
Keywords: singularity, Liouville theorem, conformal Hessian equation.

According to Caffarelli, Nirenberg, and Spruck [Caffarelli et al. 1985], we say u is *k-admissible* with respect to $\sigma_k(A^g)$ if $u \in \Gamma^k$, where Γ^k is defined by

$$\Gamma^k = \{u \in C^2(\Omega) : \sigma_s(A^g) > 0, s = 1, 2, \ldots, k\}.$$

Equation (1-1) is raised in conformal geometry and has been studied extensively. For the critical case $\alpha = 0$ of (1-1), the isolated singularities at the origin were completely understood by Caffarelli, Gidas, and Spruck for $k = 1$ [Caffarelli et al. 1989] and by Han, Li, and Teixeira for $k > 1$ [Han et al. 2010], where they employed the method of moving planes; while for the subcritical case $\alpha \in (0, k)$, the isolated singularities were classified by Gidas and Spruck for $k = 1$ [1981] and by González for $1 < k < (n-1)/2$ [2006a]. The local behavior of singularities of the conformal Hessian problems was also studied by Chang, Gursky, and Yang [Chang et al. 2003], González [2006b], and Gursky and Viacolvsky [2006].

In this paper, we bring the results of [González 2006a] to completion. The main arguments in [Gidas and Spruck 1981] and [González 2006a] are some techniques of integration by parts which were due originally to Obata [1962]. Compared with the semilinear case $k = 1$, for $k > 1$, the problems are fully nonlinear and more complicated. The "almost" divergent structure for $\sigma_k(A^g)$ explored by González [2005] allows one to carry out integration by parts for the fully nonlinear cases. We develop the arguments in [Gidas and Spruck 1981] and [González 2006a] to deal with the special case $n = 2k + 1$. Note that the special case $k = 1, n = 3$ was treated separately in [Gidas and Spruck 1981]. Of course, our main idea is to use the "almost" divergent structure for $\sigma_k(A^g)$.

Our main result reads as follows.

Theorem 1.1. *Let $\alpha \in (0, k)$, $n = 2k + 1$ and $u > 0$ be a k-admissible solution of*

$$\sigma_k(A^g) = u^\alpha \quad \text{in } B \setminus \{0\} \tag{1-3}$$

with $u^{-1} \in C^3(B \setminus \{0\})$. Then there exists a constant C such that

$$u^{-1} \le \frac{C}{|x|^{2k/(2k-\alpha)}} \quad \text{near } x = 0.$$

Furthermore, if u^{-1} is not bounded near the origin, we also get

$$u^{-1} \ge \frac{1/C}{|x|^{2k/(2k-\alpha)}} \quad \text{near } x = 0.$$

González [2006a] proved the above results for $n > 2k + 1$. The main ingredient in González's proof is the following integral estimate.

Proposition 1.2. *Let $\alpha \in (0, k)$, $n > 2k + 1$ and $u > 0$ be a k-admissible solution of (1-3). Let $r > 0$ small and $M > 0$ be such that*

$$\{r < |x| < Mr\} \subset B \setminus \{0\}.$$

Then

$$\int_{r<|x|<Mr} u^{\alpha((k+1)/k)-\delta}\, dx \le Cr^{n-(\delta-\alpha(k+1)/k)/(1-\alpha/2k)}, \tag{1-4}$$

where the constant $\delta < n + 1$ is close enough to $n + 1$ and $C > 0$ depends on M and δ but not on r.

So, to prove Theorem 1.1, we need a similar integral estimate as (1-4). In fact, in this paper, we prove the integral estimate as follows.

Proposition 1.3. *Let $\alpha \in (0, k)$, $n = 2k + 1$, and $u > 0$ be a k-admissible solution of (1-3). Let $r > 0$ small and $M > 0$ be such that*

$$\{r < |x| < Mr\} \subset B \setminus \{0\}.$$

Then

$$\int_{r<|x|<Mr} u^{\alpha(k+1)/k-n-1}\, dx \le \frac{C}{r}, \tag{1-5}$$

where the constant $C > 0$ depends on M but not on r.

By this estimate, the rest of the proof of Theorem 1.1 can be done as in [González 2006a], and we omit it in this paper.

Meanwhile, by the method shown in this paper, we are able to get the entire Liouville theorem for this special case of conformal Hessian equations. Precisely, we have the following.

Theorem 1.4. *For $\alpha \in [0, +\infty)$ and $n = 2k + 1$, consider the problem*

$$\sigma_k(A^g) = u^\alpha \quad \textit{in } \mathbb{R}^n. \tag{1-6}$$

(i) *If $\alpha > 0$, (1-6) has no positive k-admissible solution.*

(ii) *If $\alpha = 0$, any positive k-admissible solution of (1-6) must be a quadratic polynomial*

$$u = a + b|x - x_0|^2 \tag{1-7}$$

for some fixed $x_0 \in \mathbb{R}^n$ and positive constants a, b.

Li and Li [2005] classified all the solutions of (1-6) for $\alpha \in [0, +\infty)$ via the method of moving planes. But our proof of Theorem 1.4 is quite different from that in [Li and Li 2005], and similar to that in [Chang et al. 2003], where they treated the case $k = 2$.

The paper is organized as follows. In Section 2, we collect some known algebraic properties of σ_k. In Section 3, we deduce some preparation decomposition results. The proofs of Proposition 1.3 and Theorem 1.4 are given in Section 4.

2. Algebraic properties of σ_k

Throughout the paper the summation convention for repeated indices is used.

For a general $n \times n$ symmetric matrix A, consider its eigenvalues $\lambda_1, \dots, \lambda_n$ and the elementary symmetric polynomial functions

$$\sigma_k = \sum_{1 \le i_1 < \cdots < i_k \le n} \lambda_{i_1} \cdots \lambda_{i_k}. \tag{2-1}$$

For $k = 1, \dots, n$, denote the Newton tensor by

$$T^k = \sigma_k I - \sigma_{k-1} A + \cdots + (-1)^k A^k = \sigma_k I - T^{k-1} A, \tag{2-2}$$

and the traceless Newton tensor by

$$L^k = \frac{n-k}{n} \sigma_k I - T^k. \tag{2-3}$$

Here we take $\sigma_0 = 1$ and $T^0_{ij} = \delta_{ij}$.

Propositions 2.1 and 2.2 are well known (see [González 2006a] and references therein) and we omit their proofs.

Proposition 2.1. *For A and T^k and L^k as above and with the constant $C > 0$ depending only on n and s, the following hold*:

(a) $(n-k)\sigma_k = \operatorname{trace}(T^k)$.

(b) $(k+1)\sigma_{k+1} = \operatorname{trace}(AT^k)$.

(c) *If* $\sigma_1, \dots, \sigma_k > 0$, *then* T^s *is positive definite for* $s = 1, \dots, k-1$, *and hence* $\| {T_{ij}}^s \| \le C\sigma_s$.

(d) *If* $\sigma_1, \dots, \sigma_k > 0$, *then* $\sigma_s \le C(\sigma_1)^s$ *for* $s = 1, \dots, k$.

(e) *If* $\sigma_1, \dots, \sigma_k > 0$, *then* ${L_{ij}}^s {L_{ij}}^1 \ge 0$ *for* $s = 1, \dots, k$ *with equality if and only if* $L^1 = 0$.

Proposition 2.2. *For $A = A^g$, the Schouten tensor as in (1-2), and T^k and L^k defined as in (2-2) and (2-3), we have the following divergence formulas*:

(a) $\nabla^g_j {T_{ij}}^k = 0$,

(b) $\partial_j {T_{ij}}^k = -(n-k)\sigma_k u_i u^{-1} + n {T_{ij}}^k u_j u^{-1}$,

(c) $k\sigma_k = u \partial_j (u_i {T_{ij}}^{k-1}) - n {T_{ij}}^{k-1} u_i u_j + \dfrac{n-k+1}{2} \sigma_{k-1} |Du|^2$,

(d) $\partial_j {L_{ij}}^k = \dfrac{n-k}{n} \partial_i \sigma_k + n {L_{ij}}^k u_j u^{-1}$, *where ∇_j^g is the j-th covariant derivative with respect to the metric $g = u^{-2}\,dx^2$ and $\partial_i = \partial/\partial x_i$ is the usual derivative.*

3. Some decomposition results

Let $u > 0$ be in Γ^k. In the rest of the paper, we write $\sigma_s(A^g)$ simply as σ_s.

Let η be a smooth cut-off function supported in the ball B_{4r} satisfying

$$|D^m \eta| \lesssim \frac{1}{r^m}.$$

We use $\lesssim$, $\backsimeq$, etc. to drop some positive constants independent of r and u, and D^m means the usual m-th order multiple derivative.

Let δ, θ be constants which will be chosen later. For $s = 1, \dots, k$, set

$$b_s = -\frac{(n+\delta)k + (2k+\delta)s}{s!2^s}(n+\delta+1)\cdots(n+\delta+s-1)$$

and

$$B_s = \int \sigma_{k-s}|Du|^{2s}u^\delta \eta^\theta\,dx,$$

$$M_s = \int T_{ij}{}^{k-s}u_iu_j|Du|^{2(s-1)}u^\delta\eta^\theta\,dx,$$

$$E_s = \int T_{ij}{}^{k-s}u_i\eta_j|Du|^{2(s-1)}u^{\delta+1}\eta^{\theta-1}\,dx.$$

Throughout the paper, for convenience, we drop the domain in integrations; one can assume that all integrations are over a suitable domain such as $\operatorname{supp}\eta$ without confusion.

For computational convenience, we give the following recursion formula.

Lemma 3.1. *For $s = 1, \dots, k-1$,*

$$m_sM_s = m_{s+1}M_{s+1} + \frac{k+s}{2s}m_sB_s - \frac{n-k+s+1}{2(n+\delta+s+1)}m_{s+1}B_{s+1} + c_{s+1}E_{s+1}, \tag{3-1}$$

where

$$m_i = \frac{2i(n+\delta+i)}{(n+\delta)k+(2k+\delta)i}b_i$$

and

$$c_i = \theta\frac{m_i}{n+\delta+i}$$

for $i = 1, \dots, k$.

Proof. Using the above notation, by (2-2), Proposition 2.2(c), and integration by parts, we get

$$
\begin{aligned}
\text{(3-2)}\quad & m_s M_s \\
&= m_s \int T_{ij}{}^{k-s} u_i u_j |Du|^{2(s-1)} u^\delta \eta^\theta \, dx \\
&= m_s \int (\sigma_{k-s}\delta_{ij} - T_{il}{}^{k-s-1}(u u_{lj} - \tfrac{1}{2}|Du|^2 \delta_{lj})) u_i u_j |Du|^{2(s-1)} u^\delta \eta^\theta \, dx \\
&= m_s B_s + \frac{m_s}{2} M_{s+1} - \frac{m_s}{2s} \int u_i T_{il}{}^{k-s-1} \partial_l (|Du|^{2s}) u^{\delta+1} \eta^\theta \, dx \\
&= m_s B_s + \frac{m_s}{2} M_{s+1} + \frac{m_s}{2s} \int \partial_l (u_i T_{il}{}^{k-s-1}) |Du|^{2s} u^{\delta+1} \eta^\theta \, dx \\
&\qquad\qquad + \frac{m_s}{2s}(\delta+1) M_{s+1} + \theta \frac{m_s}{2s} E_{s+1} \\
&= m_s B_s + \frac{m_s}{2} M_{s+1} + \frac{m_s}{2s} \int \Big[(k-s)\sigma_{k-s} + n T_{ij}{}^{k-s-1} u_i u_j \\
&\qquad\qquad - \frac{n-k+s+1}{2} \sigma_{k-s-1} |Du|^2 \Big] |Du|^{2s} u^\delta \eta^\theta \, dx \\
&\qquad\qquad + \frac{m_s}{2s}(\delta+1) M_{s+1} + \theta \frac{m_s}{2s} E_{s+1} \\
&= m_{s+1} M_{s+1} + \frac{k+s}{2s} m_s B_s - \frac{n-k+s+1}{2(n+\delta+s+1)} m_{s+1} B_{s+1} + c_{s+1} E_{s+1}. \quad \square
\end{aligned}
$$

Now we have the decomposition for the integral for σ_k.

Proposition 3.2.

$$
\int k \sigma_k u^\delta \eta^\theta \, dx = \sum_{s=1}^{k} b_s B_s + \sum_{s=1}^{k} c_s E_s. \tag{3-3}
$$

Proof. By Proposition 2.2(c) and integration by parts we get

$$
\begin{aligned}
\text{(3-4)}\quad & \int k \sigma_k u^\delta \eta^\theta \, dx \\
&= \int \Big[u \partial_j (u_i T_{ij}{}^{k-1}) - n T_{ij}{}^{k-1} u_i u_j + \frac{n-k+1}{2} \sigma_{k-1} |Du|^2 \Big] u^\delta \eta^\theta \, dx \\
&= \frac{n-k+1}{2} \int \sigma_{k-1} |Du|^2 u^\delta \eta^\theta \, dx - n \int T_{ij}{}^{k-1} u_i u_j u^\delta \eta^\theta \, dx \\
&\qquad\qquad - \int T_{ij}{}^{k-1} u_i \partial_j (u^{\delta+1} \eta^\theta) \, dx \\
&= \frac{n-k+1}{2} \int \sigma_{k-1} |Du|^2 u^\delta \eta^\theta \, dx - \theta \int T_{ij}{}^{k-1} u_i \eta_j u^{\delta+1} \eta^{\theta-1} \, dx \\
&\qquad\qquad - (n+\delta+1) \int T_{ij}{}^{k-1} u_i u_j u^\delta \eta^\theta \, dx \\
&= \frac{n-k+1}{2} B_1 + C_1 E_1 + m_1 M_1.
\end{aligned}
$$

Using the recursion formula (3-1) in (3-4) step by step, we deduce (3-3). $\square$

For the traceless Newton tensor L^k, we also have the following decomposition.

Proposition 3.3.

$$\int L_{ij}{}^k L_{ij}{}^1 u^\delta \eta^\theta\,dx \tag{3-5}$$
$$= -\frac{n-k}{n}\int \partial_i(\sigma_k)u_i u^{\delta+1}\eta^\theta\,dx - (n+1+\delta)\int L_{ij}{}^k u_i u_j u^\delta \eta^\theta\,dx$$
$$+\frac{n-k}{n(n+2+\delta)}\int \partial_i(\sigma_k)\partial_i(\eta^\theta)u^{\delta+2}\,dx - \frac{k}{n(n+2+\delta)}\int \sigma_k\Delta(\eta^\theta)u^{\delta+2}\,dx$$
$$-\frac{1}{2(n+2+\delta)}\int T_{ij}{}^{k-1}\partial_{ij}(\eta^\theta)|Du|^2u^{\delta+2}\,dx + \frac{n-k+1}{n+2+\delta}\int \sigma_{k-1}u_iu_j\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx$$
$$-\frac{n+3+\delta}{n+2+\delta}\int T_{il}{}^{k-1}u_lu_j\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx - \frac{1}{n+2+\delta}\int T_{il}{}^{k-1}u_j\partial_{ijl}(\eta^\theta)u^{\delta+3}\,dx.$$

Proof. By Proposition 2.2(d) and integration by parts we get

$$\int L_{ij}{}^k L_{ij}{}^1 u^\delta \eta^\theta\,dx \tag{3-6}$$
$$= \int L_{ij}{}^k u_{ij} u^{\delta+1}\eta^\theta\,dx$$
$$= -\int \partial_j(L_{ij}{}^k)u_iu^{\delta+1}\eta^\theta\,dx - (\delta+1)\int L_{ij}{}^k u_iu_ju^\delta\eta^\theta\,dx - \int L_{ij}{}^k u_i\partial_j(\eta^\theta)u^{\delta+1}\,dx$$
$$= -\int \Big[\frac{n-k}{n}\partial_i(\sigma_k) + nL_{ij}{}^k u_ju^{-1}\Big]u_iu^{\delta+1}\eta^\theta\,dx$$
$$-(\delta+1)\int L_{ij}{}^k u_iu_ju^\delta\eta^\theta\,dx - \int L_{ij}{}^k u_i\partial_j(\eta^\theta)u^{\delta+1}\,dx$$
$$= -\frac{n-k}{n}\int \partial_i(\sigma_k)u_iu^{\delta+1}\eta^\theta\,dx - (n+\delta+1)\int L_{ij}{}^k u_iu_ju^\delta\eta^\theta\,dx$$
$$-\int L_{ij}{}^k u_i\partial_j(\eta^\theta)u^{\delta+1}\,dx.$$

For the last term in (3-6), integrating once again, we have

$$-\int L_{ij}{}^k u_i\partial_j(\eta^\theta)u^{\delta+1}\,dx \tag{3-7}$$
$$= \int \partial_i(L_{ij}{}^k)\partial_j(\eta^\theta)u^{\delta+2}dx + \int L_{ij}{}^k\partial_{ij}(\eta^\theta)u^{\delta+2}dx + (\delta+1)\int L_{ij}{}^k\partial_j(\eta^\theta)u_iu^{\delta+1}dx$$
$$= \int \Big[\frac{n-k}{n}\partial_i(\sigma_k) + nL_{ij}{}^k u_ju^{-1}\Big]\partial_i(\eta^\theta)u^{\delta+2}dx$$
$$+\int L_{ij}{}^k\partial_{ij}(\eta^\theta)u^{\delta+2}dx + (\delta+1)\int L_{ij}{}^k\partial_j(\eta^\theta)u_iu^{\delta+1}dx$$
$$= \frac{n-k}{n}\int \partial_i(\sigma_k)\partial_i(\eta^\theta)u^{\delta+2}dx + \int L_{ij}{}^k\partial_{ij}(\eta^\theta)u^{\delta+2}dx$$
$$+(n+\delta+1)\int L_{ij}{}^k\partial_j(\eta^\theta)u_iu^{\delta+1}\,dx.$$

Transposition of the term implies

$$-\int L_{ij}{}^{k}u_i\partial_j(\eta^\theta)u^{\delta+1}\,dx = \frac{n-k}{n(n+2+\delta)}\int \partial_i(\sigma_k)\partial_i(\eta^\theta)u^{\delta+2}\,dx + \frac{1}{n+2+\delta}\int L_{ij}{}^{k}\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx. \tag{3-8}$$

For the last term in (3-8), we have

$$\begin{aligned}\int L_{ij}{}^{k}\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx &= \int \Bigl(T_{il}{}^{k-1}A_{lj} - \frac{k}{n}\sigma_k\delta_{ij}\Bigr)\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx \\ &= \int T_{il}{}^{k-1}(uu_{lj} - \tfrac{1}{2}|Du|^2\delta_{lj})\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx - \frac{k}{n}\int \sigma_k\Delta(\eta^\theta)u^{\delta+2}\,dx \\ &= -\frac{k}{n}\int \sigma_k\Delta(\eta^\theta)u^{\delta+2}\,dx - \tfrac{1}{2}\int T_{ij}{}^{k-1}\partial_{ij}(\eta^\theta)|Du|^2u^{\delta+2}\,dx \\ &\qquad + \int T_{il}{}^{k-1}u_{lj}\partial_{ij}(\eta^\theta)u^{\delta+3}\,dx.\end{aligned} \tag{3-9}$$

For the last term in (3-9), by Proposition 2.2(b), we compute

$$\begin{aligned}\int T_{il}{}^{k-1}u_{lj}\partial_{ij}(\eta^\theta)u^{\delta+3}\,dx &= -\int \partial_l(T_{il}{}^{k-1})u_j\partial_{ij}(\eta^\theta)u^{\delta+3}\,dx - \int T_{il}{}^{k-1}u_j\partial_{ijl}(\eta^\theta)u^{\delta+3}\,dx \\ &\qquad - (\delta+3)\int T_{il}{}^{k-1}u_ju_l\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx \\ &= -\int [-(n-k+1)\sigma_{k-1}u_iu^{-1} + nT_{il}{}^{k-1}u_lu^{-1}]u_j\partial_{ij}(\eta^\theta)u^{\delta+3}\,dx \\ &\qquad - \int T_{il}{}^{k-1}u_j\partial_{ijl}(\eta^\theta)u^{\delta+3}\,dx - (\delta+3)\int T_{il}{}^{k-1}u_ju_l\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx \\ &= (n-k+1)\int \sigma_{k-1}u_iu_j\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx - \int T_{il}{}^{k-1}u_j\partial_{ijl}(\eta^\theta)u^{\delta+3}\,dx \\ &\qquad - (n+\delta+3)\int T_{il}{}^{k-1}u_ju_l\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx.\end{aligned} \tag{3-10}$$

Inserting this into (3-9), we get

$$\begin{aligned}\int L_{ij}{}^{k}\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx &= -\frac{k}{n}\int \sigma_k\Delta(\eta^\theta)u^{\delta+2}\,dx - \tfrac{1}{2}\int T_{ij}{}^{k-1}\partial_{ij}(\eta^\theta)|Du|^2u^{\delta+2}\,dx \\ &\qquad + (n-k+1)\int \sigma_{k-1}u_iu_j\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx \\ &\qquad - (n+3+\delta)\int T_{il}{}^{k-1}u_ju_l\partial_{ij}(\eta^\theta)u^{\delta+2}\,dx - \int T_{il}{}^{k-1}u_j\partial_{ijl}(\eta^\theta)u^{\delta+3}\,dx.\end{aligned} \tag{3-11}$$

Substituting this into (3-8) and then (3-6), we get (3-5) as desired. □

To end this section, we give the estimate on the "error" terms "E_s" in (3-3).

Lemma 3.4.

$$|E_s| \lesssim \varepsilon \sum_{m=s}^{k} B_m + \frac{1}{r^{2k}} \int u^{\delta+2k} \eta^{\theta-2k}\, dx. \tag{3-12}$$

Proof. First, by $|D\eta| \lesssim 1/r$ and Proposition 2.1(c), we have

$$|E_s| \lesssim \frac{1}{r} \int \sigma_{k-s} |Du|^{2s-1} u^{\delta+1} \eta^{\theta-1}\, dx.$$

Using Young's inequality with exponent pair $(2s/(2s-1), 2s)$ and $\varepsilon > 0$ small, the last inequality turns into

$$|E_s| \lesssim \varepsilon \int \sigma_{k-s} |Du|^{2s} u^{\delta} \eta^{\theta}\, dx + \frac{C(\varepsilon)}{r^{2s}} \int \sigma_{k-s} u^{\delta+2s} \eta^{\theta-2s}\, dx. \tag{3-13}$$

For the last term of (3-13), by Proposition 2.2(c), we deduce

$$\begin{aligned}
&\frac{C(\varepsilon)}{r^{2s}} \int \sigma_{k-s} u^{\delta+2s} \eta^{\theta-2s}\, dx \\
&\backsimeq \frac{1}{r^{2s}} \int \Big[u \partial_j (u_i {T_{ij}}^{k-s-1}) - n {T_{ij}}^{k-s-1} u_i u_j \\
&\qquad\qquad + \frac{n-k+s+1}{2} \sigma_{k-s-1} |Du|^2 \Big] u^{\delta+2s} \eta^{\theta-2s}\, dx \\
&\backsimeq \frac{1}{r^{2s}} \int \sigma_{k-s-1} |Du|^2 u^{\delta+2s} \eta^{\theta-2s}\, dx - \frac{1}{r^{2s}} \int {T_{ij}}^{k-s-1} u_i u_j u^{\delta+2s} \eta^{\theta-2s}\, dx \\
&\qquad\qquad - \frac{1}{r^{2s}} \int {T_{ij}}^{k-s-1} u_i \eta_j u^{\delta+2s+1} \eta^{\theta-2s-1}\, dx \\
&\lesssim \frac{1}{r^{2s}} \int \sigma_{k-s-1} |Du|^2 u^{\delta+2s} \eta^{\theta-2s}\, dx + \frac{1}{r^{2s+1}} \int \sigma_{k-s-1} |Du| u^{\delta+2s+1} \eta^{\theta-2s-1}\, dx \\
&\lesssim \varepsilon \int \sigma_{k-s-1} |Du|^{2(s+1)} u^{\delta} \eta^{\theta}\, dx + \frac{C(\varepsilon)}{r^{2(s+1)}} \int \sigma_{k-s-1} u^{\delta+2(s+1)} \eta^{\theta-2(s+1)}\, dx,
\end{aligned} \tag{3-14}$$

where we have used Young's inequality in the last step in (3-13).

Substituting (3-14) into (3-13) step by step shows (3-12). □

4. Proofs of Proposition 1.3 and Theorem 1.4

For $n = 2k+1$, if we choose $\delta = -2k = 1-n$, (3-12) implies

$$|E_s| \lesssim \varepsilon \sum_{m=s}^{k} B_m + r. \tag{4-1}$$

Moreover, by this choice of δ we see that $b_s < 0 (s = 1, 2, \ldots, k)$. Hence if we take ε small enough, combining (3-3) with (4-1), we have

$$\int \sigma_k u^{1-n} \eta^\theta \, dx + \sum_{s=1}^{k} B_s \lesssim r. \tag{4-2}$$

On the other hand, if we choose $\delta = -n-1$ in (3-5) , then

$$\begin{aligned}
&\int L_{ij}{}^k L_{ij}{}^1 u^{-n-1} \eta^\theta \, dx \\
&= -\frac{n-k}{n} \int \partial_i(\sigma_k) u_i u^{-n} \eta^\theta \, dx + \frac{n-k}{n} \int \partial_i(\sigma_k) \partial_i(\eta^\theta) u^{1-n} \, dx - \frac{k}{n} \int \sigma_k \Delta(\eta^\theta) u^{1-n} \, dx \\
&\quad - \tfrac{1}{2} \int T_{ij}{}^{k-1} \partial_{ij}(\eta^\theta) |Du|^2 u^{1-n} \, dx + (n-k+1) \int \sigma^{k-1} u_i u_j \partial_{ij}(\eta^\theta) u^{1-n} \, dx \\
&\quad - 2 \int T_{il}{}^{k-1} u_l u_j \partial_{ij}(\eta^\theta) u^{1-n} \, dx - \int T_{il}{}^{k-1} u_j \partial_{ijl}(\eta^\theta) u^{2-n} \, dx.
\end{aligned} \tag{4-3}$$

By (1-1) and $|D^m \eta| \lesssim 1/r^m$ we deduce

$$\begin{aligned}
&\int L_{ij}{}^k L_{ij}{}^1 u^{-n-1} \eta^\theta \, dx \\
&\quad \lesssim -\frac{n-k}{n} \alpha \int |Du|^2 u^{\alpha-n-1} \eta^\theta \, dx + \frac{n-k}{n} \alpha\theta \int u_i \eta_i u^{\alpha-n} \eta^{\theta-1} \, dx \\
&\qquad + \frac{1}{r^2} \int u^{\alpha+1-n} \eta^{\theta-2} \, dx + \frac{1}{r^2} \int \sigma_{k-1} |Du|^2 u^{1-n} \eta^{\theta-2} \, dx \\
&\qquad + \frac{1}{r^3} \int \sigma_{k-1} |Du| u^{2-n} \eta^{\theta-3} \, dx.
\end{aligned} \tag{4-4}$$

Using Young's inequality, by (4-4), we can get

$$\begin{aligned}
&\int L_{ij}{}^k L_{ij}{}^1 u^{-n-1} \eta^\theta \, dx \\
&\quad \lesssim \Big(\varepsilon - \frac{n-k}{n}\Big) \alpha \int |Du|^2 u^{\alpha-n-1} \eta^\theta \, dx + \frac{1}{r^2} \int u^{\alpha+1-n} \eta^{\theta-2} \, dx \\
&\qquad + \frac{1}{r^2} \int \sigma_{k-1} |Du|^2 u^{1-n} \eta^{\theta-2} \, dx + \frac{1}{r^4} \int \sigma_{k-1} u^{3-n} \eta^{\theta-4} \, dx.
\end{aligned} \tag{4-5}$$

For the last term of (4-5), using (3-14) (with $\delta = 1-n$) step by step, we have

$$\begin{aligned}
\frac{1}{r^4} \int \sigma_{k-1} u^{3-n} \eta^{\theta-4} \, dx &\lesssim \frac{1}{r^2} \Big[\sum_{s=2}^{k} B_s + \frac{1}{r^{2k}} \int u^{1-n+2k} \eta^{\theta-2-2k} \, dx \Big] \\
&\lesssim \frac{1}{r^2} \sum_{s=2}^{k} B_s + \frac{1}{r}.
\end{aligned} \tag{4-6}$$

Taking ε small, inserting (4-6) into (4-5), and combining with (4-2) (replacing θ with $\theta-2$), we get

$$\text{(4-7)}\quad \int L_{ij}{}^{k}L_{ij}{}^{1}u^{-n-1}\eta^{\theta}\,dx+\alpha\int |Du|^2u^{\alpha-n-1}\eta^{\theta}\,dx \lesssim \frac{1}{r^2}\left[\int u^{\alpha+1-n}\eta^{\theta-2}\,dx+\sum_{s=1}^{k}B_s\right]+\frac{1}{r}\lesssim\frac{1}{r}.$$

Now, from (4-7), we can prove Theorem 1.4 and Proposition 1.3.

Proof of Theorem 1.4. Let $\eta\equiv 1$ in B_r, $0<\eta<1$ in $B_{2r}\setminus B_r$. Taking $r\to+\infty$ in (4-7), we can get

$$\text{(4-8)}\quad \int_{\mathbb{R}^n} L_{ij}{}^{k}L_{ij}{}^{1}u^{-n-1}\,dx+\alpha\int_{\mathbb{R}^n}|Du|^2u^{\alpha-n-1}\,dx\le 0.$$

By Proposition 2.1(e), if $\alpha>0$, (4-8) shows u must be a positive constant solution of (1-6), which is impossible; if $\alpha=0$, (4-8) shows $L^1=0$ and hence u must be the quadratic polynomial as in (1-7). □

Proof of Proposition 1.3. Let $\eta\equiv 1$ for $r\le|x|\le Mr$ and $\eta=0$ for $0<|x|<r/2$, $2Mr<|x|$. By (1-3) and Proposition 2.1(d) we have

$$\text{(4-9)}\quad \int u^{\alpha/k+\alpha-n-1}\eta^{\theta}\,dx=\int(\sigma_k)^{1/k}u^{\alpha-n-1}\eta^{\theta}\,dx\lesssim\int\sigma_1u^{\alpha-n-1}\eta^{\theta}\,dx = -\frac{n}{2}\int|Du|^2u^{\alpha-n-1}\eta^{\theta}\,dx+\int\Delta u u^{\alpha-n}\eta^{\theta}\,dx.$$

For the last term in (4-9), integrating by parts and using Young's inequality, we deduce

$$\text{(4-10)}\quad \int\Delta u u^{\alpha-n}\eta^{\theta}\,dx=(n-\alpha)\int|Du|^2u^{\alpha-n-1}\eta^{\theta}\,dx-\theta\int u_i\eta_iu^{\alpha-n}\eta^{\theta-1}dx \lesssim(n-\alpha+\varepsilon)\int|Du|^2u^{\alpha-n-1}\eta^{\theta}\,dx+\frac{1}{r^2}\int u^{\alpha-n+1}\eta^{\theta-2}dx.$$

Inserting this into (4-9) and combining with (4-7)and (4-2), we have

$$\text{(4-11)}\quad \int u^{((k+1)/k)\alpha-n-1}\eta^{\theta}\,dx \lesssim\left(\frac{n}{2}-\alpha+\varepsilon\right)\int|Du|^2u^{\alpha-n-1}\eta^{\theta}\,dx+\frac{1}{r^2}\int u^{\alpha-n+1}\eta^{\theta-2}\,dx\lesssim\frac{1}{r}.$$

This implies (1-5) and hence the proof of Proposition 1.3 is completed. □

Acknowledgments

The author thanks Professor Xi-Nan Ma for constant encouragement and useful discussions. He also thanks the referee for a careful reading and useful suggestions.

References

[Caffarelli et al. 1985] L. Caffarelli, L. Nirenberg, and J. Spruck, "The Dirichlet problem for nonlinear second-order elliptic equations, III: Functions of the eigenvalues of the Hessian", *Acta Math.* **155**:3-4 (1985), 261–301. MR 87f:35098 Zbl 0654.35031

[Caffarelli et al. 1989] L. A. Caffarelli, B. Gidas, and J. Spruck, "Asymptotic symmetry and local behavior of semilinear elliptic equations with critical Sobolev growth", *Comm. Pure Appl. Math.* **42**:3 (1989), 271–297. MR 90c:35075 Zbl 0702.35085

[Chang et al. 2003] S.-Y. A. Chang, M. J. Gursky, and P. C. Yang, "Entire solutions of a fully nonlinear equation", pp. 43–60 in *Lectures on partial differential equations*, edited by S.-Y. A. Chang et al., New Stud. Adv. Math. **2**, Int. Press, Somerville, MA, 2003. MR 2005b:53053 Zbl 1183.53035

[Gidas and Spruck 1981] B. Gidas and J. Spruck, "Global and local behavior of positive solutions of nonlinear elliptic equations", *Comm. Pure Appl. Math.* **34**:4 (1981), 525–598. MR 83f:35045 Zbl 0465.35003

[González 2005] M. d. M. González, "Singular sets of a class of locally conformally flat manifolds", *Duke Math. J.* **129**:3 (2005), 551–572. MR 2006d:53034 Zbl 1088.53023

[González 2006a] M. d. M. González, "Classification of singularities for a subcritical fully nonlinear problem", *Pacific J. Math.* **226**:1 (2006), 83–102. MR 2007h:35091 Zbl 1192.35050

[González 2006b] M. d. M. González, "Removability of singularities for a class of fully non-linear elliptic equations", *Calc. Var. Partial Differential Equations* **27**:4 (2006), 439–466. MR 2007g:35052 Zbl 1151.35347

[Gursky and Viaclovsky 2006] M. J. Gursky and J. Viaclovsky, "Convexity and singularities of curvature equations in conformal geometry", *Int. Math. Res. Not.* **2006**:Art. ID 9689 (2006), 1–43. MR 2007b:53080 Zbl 1132.53020

[Han et al. 2010] Z.-C. Han, Y. Li, and E. V. Teixeira, "Asymptotic behavior of solutions to the σ_k-Yamabe equation near isolated singularities", *Invent. Math.* **182**:3 (2010), 635–684. MR 2011i:53045 Zbl 1211.53064

[Li and Li 2005] A. Li and Y. Y. Li, "On some conformally invariant fully nonlinear equations, II: Liouville, Harnack and Yamabe", *Acta Math.* **195** (2005), 117–154. MR 2007d:53053 Zbl 1216.35038

[Obata 1962] M. Obata, "Certain conditions for a Riemannian manifold to be isometric with a sphere", *J. Math. Soc. Japan* **14** (1962), 333–340. MR 25 #5479 Zbl 0115.39302

Received August 26, 2012. Revised December 30, 2012.

QIANZHONG OU
DEPARTMENT OF MATHEMATICS
HEZHOU UNIVERSITY
HEZHOU, 542800
GUANGXI PROVINCE
CHINA
ouqzh@163.com

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.129

ATTACHING HANDLES TO DELAUNAY NODOIDS

FRANK PACARD AND HAROLD ROSENBERG

For all $m \in \mathbb{N} - \{0\}$, we prove the existence of a one-dimensional family of genus m, constant mean curvature (equal to 1) surfaces which are complete, immersed in $\mathbb{R}^3$, and have two Delaunay ends asymptotic to nodoidal ends. Moreover, these surfaces are invariant under the group of isometries of $\mathbb{R}^3$ leaving a horizontal regular polygon with $m+1$ sides fixed.

1. Introduction

Delaunay surfaces are complete, noncompact constant mean curvature surfaces of revolution in $\mathbb{R}^3$ which are either embedded or immersed. The embedded Delaunay surfaces are usually referred to as *unduloids*. The elements of this family are generated by *roulettes* of ellipses [Eells 1987], and they interpolate between a right cylinder $S^1\left(\frac{1}{2}\right) \times \mathbb{R} \subset \mathbb{R}^3$ and a singular surface which is constituted by infinitely many tangent spheres of radius 1 which are periodically arranged along the vertical axis. Close to the singular limit, the Delaunay unduloids can be understood as infinitely many spheres of radius 1 which are disjoint, arranged periodically along the vertical axis, each sphere being connected to its two nearest neighbors by catenoids, whose rotational axis is the vertical axis, which have been scaled by a small factor $\tau > 0$.

The immersed Delaunay surfaces are referred to as *nodoids*. The elements of this family are generated by *roulettes* of hyperbolas [Eells 1987]. Again, part of this family converges to infinitely many spheres of radius 1 which are periodically arranged along the vertical axis. In contrast to unduloids, close to the singular limit, the Delaunay nodoids can be understood as infinitely many spheres of radius 1 which are either disjoint or slightly overlapping and which are arranged periodically along the vertical axis, each sphere being connected to its two nearest neighbors (with which it shares a slight overlap) by catenoids that have vertical axes and that have been scaled by a small factor $\tau > 0$.

In this paper, we prove the existence of constant mean curvature surfaces which have two Delaunay ends (of nodoid type) and finite genus.

The first author is partially supported by the ANR-11-IS01-0002 grant.
MSC2010: primary 53C42; secondary 53A10.
Keywords: constant mean curvature surfaces, Delaunay surfaces.

Theorem 1.1. *For all $m \geq 1$, there exists a one parameter family of genus m constant mean curvature* (*with mean curvature equal to* 1) *surfaces which are invariant under the action of the full dihedral group* $\mathrm{Dih}^{(3)}_{m+1}$ (*the group of isometries of $\mathbb{R}^3$ leaving a horizontal regular polygon with $m+1$ sides fixed*) *and which have two Delaunay ends asymptotic to nodoidal ends.*

Let us briefly describe how these surfaces are constructed, since this will provide an opportunity to give a precise picture of the surfaces themselves.

As already mentioned, close to the singular limit, the Delaunay nodoids can be understood as infinitely many spheres of radius 1 which are either disjoint or slightly overlapping, arranged periodically along the vertical axis and which are connected together by catenoids, which have vertical axes, and which are scaled by a small factor $\tau > 0$; these latter ones are called *catenoidal necks*. The spheres of radius 1 arranged along the vertical axis can be ordered (by the height of their centers) and can be indexed by $j \in \mathbb{Z}$ (without loss of generality, we can assume that the center of the sphere of index j is at height $2j+1$). In this description, one can check that the distance between the centers of two consecutive spheres can be expanded as

$$d_\tau = 2 + 2\tau \log \tau + \mathbb{O}(\tau),$$

as τ tends to 0. In order to obtain the surfaces of Theorem 1.1, instead of connecting the sphere indexed by 0 and the sphere indexed by 1 using *one* catenoidal neck, we connect these two spheres using $m+1$ catenoids which are scaled by a factor

$$\tilde{\tau} = \frac{\tau}{m+1} + \mathbb{O}(\tau^{3/2}),$$

and whose axes are vertical and pass through the vertices of a horizontal regular polygon (with $m+1$ sides) of size $\rho > 0$. We show that this construction is successful provided the parameter ρ, which measures the size of the polygon, is carefully chosen (as a function of τ), and, in fact, we find that

$$\rho^2 = \frac{m}{m+1}\frac{\tau}{2} + \mathbb{O}(\tau^{5/4}).$$

Notice that all the surfaces we construct have the same small vertical flux (we refer to Section 4 for a definition of the flux of a Delaunay surface).

Our construction is quite flexible and provides many other interesting constant mean curvature surfaces. For example, using similar ideas and proofs, one can also construct singly periodic constant mean curvature surfaces with (infinite) topology: starting with the spheres of radius 1 which are periodically arranged along the vertical axis and which are either disjoint or slightly overlapping, we can choose to connect *any* two consecutive spheres using $m+1$ catenoids scaled by a factor $\tilde{\tau}$ whose axes are vertical and pass through the vertices of a horizontal regular polygon (with

$m+1$ sides) of size $\rho > 0$. More generally, there is strong evidence for the following.

It should be possible to construct constant mean curvature surfaces starting from a subset $\mathfrak{Z} \subset \mathbb{Z}$ and assuming that, for all $j \in \mathbb{Z} - \mathfrak{Z}$, we decide to connect the sphere of index j to the sphere of index $j+1$ using *one* catenoid whose axis is the vertical axis and which is scaled by a factor τ, while, when $j \in \mathfrak{Z}$, we decide to connect the sphere of index j to the sphere of index $j+1$ using $m+1$ catenoids whose axes are vertical and pass through the vertices of a small horizontal regular polygon (with $m+1$ sides) of size $\rho > 0$ with

$$\rho^2 \sim \frac{m}{m+1}\,\frac{\tau}{2},$$

and which are scaled by a factor $\tilde{\tau} \sim \tau/(m+1)$. We believe that this configuration can be perturbed into a genuine constant mean curvature surface.

We mention that the present construction is very much inspired by [Hauswirth and Pacard 2007], where the authors perform a construction of minimal surfaces in $\mathbb{R}^3$ that have finite genus and two Riemann type ends. In fact, part of the analysis in the present paper parallels the analysis in [Hauswirth and Pacard 2007]. Nevertheless, in the present situation, some extra technical difficulties arise in the construction (see Section 6), since the points where the connected sum is performed are located at the vertices of a polygon whose size tends to 0 as the parameter τ tends to 0.

We end the introduction by giving an overview of the paper. In Section 2 we recall some well known facts about the mean curvature operator of normal graphs with special emphasize on the differential of the mean curvature operator. Section 3 is concerned with harmonic extensions on half cylinders, for which we prove some decay properties. Section 4 is quite long. It starts with a careful description of the Delaunay nodoids as the Delaunay parameter τ tends to 0 (that is, close to the singular limit). Then we proceed with the analysis of the Jacobi operator about a Delaunay surface as the Delaunay parameter tends to 0. Finally, starting on page 151, we apply the implicit function theorem about a half nodoid (which is a constant mean curvature surface with one boundary and one Delaunay end) to prove the existence of an infinite-dimensional family of constant mean curvature surfaces which have one Delaunay end and one boundary. These surfaces are close to the half nodoid we started with and are parametrized by their boundary data. In Section 6, we perform a similar analysis starting from the catenoid. As a result, we obtain the existence of an infinite-dimensional family of constant mean curvature surfaces which have two boundaries, are close to a truncated catenoid, and are parametrized by their boundary data. In Section 6, we start with a unit sphere, from which we excise one small disc close to the north pole and $m+1$ small discs arranged symmetrically at the vertices of a regular polygon near the south pole. We perturb this surface with $m+2$ boundaries applying the implicit function theorem to obtain an infinite-dimensional family of constant mean curvature surfaces which are parametrized by their boundary data. In the final section, we explain how all these pieces can be connected together to produce the surfaces in Theorem 1.1. At this stage, the problem then reduces to being able to choose the boundary data of the different summands so that their union is a $\mathscr{C}^1$ surface, since elliptic regularity theory will imply that what we have built is a smooth surface of constant mean curvature.

The construction relies heavily on the analysis of elliptic operators on noncompact spaces as in [Melrose 1993; Mazzeo 1991; Lockhart and McOwen 1985]. It is true that similar techniques and ideas have already been used in many constructions, but the proofs are usually hard to read for nonspecialists, since they always refer to results which are difficult to find in the literature in the precise form they are needed. This is why we have decided to present complete proofs based on simple well-known tools; we hope that this will help the interested reader master these techniques.

Finally, we mention a problem related to our work. To introduce this problem, we consider Σ to be the union of the upper hemisphere of the sphere of radius 1 centered at the points $(0, 0, -1)$ and the lower hemisphere of the sphere of radius 1 centered at the points $(0, 0, 1)$. The existence of unduloids, nodoids with small Delaunay parameters, and the surfaces we construct in this paper shows that, for all $\epsilon > 0$, there exist infinitely many surfaces of constant mean curvature 1 that are included in an ϵ-tubular neighborhood of the unduloid and are not congruent. Obviously a similar result holds for the surface Σ.

Now, if we consider two radius 1 spheres tangent at a point, can we find constant mean curvature ($= 1$) surfaces (with no boundary) in any small tubular neighborhood of this configuration? In fact, we can not even answer the following (apparently) simpler but striking question: is there any compact mean curvature ($= 1$) surface (with no boundary) near a radius 1 sphere? More precisely, is there an $\epsilon_0 > 0$ such that if Σ is a mean curvature ($= 1$) surface in the ϵ_0-tubular neighborhood of a radius 1 sphere, is Σ congruent to the sphere? In other words, what is the *form* of a compact constant mean curvature surface?

2. Generalities

The mean curvature. We gather some basic material concerning the mean curvature of a surface in Euclidean space. All these results are well known, but we feel that collecting them here makes the paper easier to read. This also gives us an opportunity to introduce some of the notation we use throughout the paper. We refer to [Colding and Minicozzi 2011; Lawson 1977] for further details.

Let us assume that Σ is a surface embedded in $\mathbb{R}^3$. We denote by g the metric induced on Σ by the Euclidean metric $\mathring{g}$, and by h the second fundamental form, defined by

$$h(t_1, t_2) = -\mathring{g}(\nabla_{t_1} N, t_2)$$

for all $t_1, t_2 \in T\Sigma$. Here N is a unit normal vector field on Σ. In this paper, we agree that the mean curvature of a surface is defined to be the *average* of the principal curvatures, or, since we are interested in 2-dimensional surfaces, the half of the trace of the second fundamental form. Hence the mean curvature of Σ is given by

$$H := \tfrac{1}{2}\operatorname{tr}^g h,$$

and the mean curvature vector is then given by $\vec{H} := HN$.

For computational purposes, we recall that the mean curvature appears in the first variation of the area functional. More precisely, given w, a sufficiently small smooth function which is defined on Σ and has compact support, we consider the surface Σ_w, which is the normal graph over Σ for the function w. Namely,

$$\Sigma \ni p \mapsto p + w(p)N(p) \in \Sigma_w.$$

We denote by A_w the area of the surface Σ_w (we assume that this area is finite). Then

$$DA_{|w=0}(v) = -2\int_\Sigma H v \, d\mathrm{vol}_g.$$

In the case where surfaces close to Σ are parametrized as graphs over Σ using a vector field $\tilde{N}$ which is transverse to Σ but which is not necessarily a unit normal vector field, the previous formula has to be modified. Let us denote by $\tilde{\Sigma}_w$ the surface which is the graph over Σ, using the vector field $\tilde{N}$, for some sufficiently small smooth function w. Namely,

$$\Sigma \ni p \mapsto p + w(p)\tilde{N}(p) \in \tilde{\Sigma}_w.$$

We denote by $\tilde{A}_w$ the area of this surface. The previous formula must be changed to

$$DA\tilde{A}_{|w=0}(v) = -2\int_\Sigma (\vec{H} \cdot \tilde{N}) v \, d\mathrm{vol}_g. \tag{2-1}$$

In the next result, we give the expression of the mean curvature H_w of the surface Σ_w in terms of w. Some notation is needed. For $z \in \mathbb{R}$ small enough, we define g_z to be the induced metric on the parallel surface

$$\Sigma_z := \Sigma + zN.$$

It is given explicitly by

$$g_z = g - 2zh + z^2 k,$$

where the tensor k is defined by

$$k(t_1, t_2) := g(\nabla_{t_1} N, \nabla_{t_2} N)$$

for all $t_1, t_2 \in T\Sigma$.

Proposition 2.1. *The mean curvature H_w of the surface Σ_w is given by the formula*

$$H_w = \tfrac{1}{2}\sqrt{1+|\nabla^{g_w} w|^2}\,\mathrm{tr}^{g_w}(h-wk) + \tfrac{1}{2}\,\mathrm{div}_{g_w}\left(\frac{\nabla^{g_w} w}{\sqrt{1+|\nabla^{g_w} w|^2}}\right) - \tfrac{1}{2}\frac{1}{\sqrt{1+|\nabla^{g_w} w|}}(h-wk)(\nabla^{g_w} w, \nabla^{g_w} w).$$

Proof. The induced metric $\tilde{g}$ on Σ_w is given by

$$\tilde{g} = g_{z=w} + dw \otimes dw.$$

This implies that

$$\det \tilde{g} = (1+|\nabla^{g_w} w|^2)\det g_w.$$

We can now compute the area of Σ_w,

$$A_w = \int_\Sigma \sqrt{1+|\nabla^{g_w} w|^2}\, d\mathrm{vol}_{g_w},$$

as well as the differential of this functional with respect to w. In doing so, one should be careful that the function w appears implicitly in the definition of g_w. Using integration by parts, we find

$$DA_w(v) = -\int_\Sigma \mathrm{div}_{g_w}\left(\frac{\nabla^{g_w} w}{\sqrt{1+|\nabla^{g_w} w|^2}}\right) v\, d\mathrm{vol}_{g_w} - \frac{1}{2}\int_\Sigma \frac{g'_w(\nabla^{g_w} w, \nabla^{g_w} w)}{\sqrt{1+|\nabla^{g_w} w|^2}} v\, d\mathrm{vol}_{g_w} + \frac{1}{2}\int_\Sigma \sqrt{1+|\nabla^{g_w} w|^2}\,\mathrm{tr}^{g_w} g'_w v\, d\mathrm{vol}_{g_w},$$

where $g'_w := \partial_z g_{z|z=w} = -2(h-wk)$. To proceed, observe that, if N_w denotes the unit normal vector field about Σ_w, we have

$$N_w = \frac{1}{\sqrt{1+|\nabla^{g_w} w|^2}}(N - \nabla^{g_w} w),$$

and hence we get

$$d\mathrm{vol}_{g_w} = (N_w \cdot N)\, d\mathrm{vol}_{\tilde{g}}.$$

The result then follows at once from (2-1). □

Linearized mean curvature operators. Again, the material in this section is well known, and we refer to [Colding and Minicozzi 2011; Lawson 1977] for a more detailed description. The Jacobi operator appears in the linearization of the mean curvature operator when nearby surfaces are parametrized as normal graphs over a given surface. Indeed, we can consider the nonlinear operator $w \mapsto H_w$, which is defined, for example, from the space $\mathscr{C}^2_{\mathrm{loc}}(\Sigma)$ into the space $\mathscr{C}^0_{\mathrm{loc}}(\Sigma)$, and it follows from Proposition 2.1 that the differential of this operator with respect to w, computed at $w = 0$, is given by

$$J := DH_{w=0} = \tfrac{1}{2}(\Delta_g + \operatorname{tr}^g k),$$

where Δ_g is the Laplace–Beltrami operator on Σ and $\operatorname{tr}^g k$ is the square of the norm of the shape operator.

Finally, we recall that if Σ is a constant mean curvature surface and if Ξ is a killing vector field (that is, Ξ generates a one parameter family of isometries), the function $N \cdot \Xi$, which is usually referred to as a *Jacobi field*, satisfies

$$J(N \cdot \Xi) = 0.$$

This is probably a good time to recall some elementary facts concerning linearized mean curvature operators when different vector fields are used. As above, we assume that we are given a vector field $\tilde{N}$ which is transverse to Σ, but which is not necessarily a unit normal vector field. Any surface close enough to Σ can be considered either as a *normal* graph over Σ or as a graph over Σ, using the vector field $\tilde{N}$. Hence we can define two nonlinear operators

$$w \mapsto H_w \quad \text{and} \quad w \mapsto \tilde{H}_w,$$

which are (respectively) the mean curvature of the *normal* graph of w and the mean curvature of the graph of w using the vector field $\tilde{N}$. The following result gives the relation between the differentials of these two operators at $w = 0$.

Proposition 2.2 [Mazzeo et al. 2001]. *The relation*

$$D\tilde{H}_{|w=0}(v) = DH_{|w=0}((\tilde{N} \cdot N)v) + (\nabla H \cdot \tilde{N})v$$

holds for any $v \in \mathscr{C}^2_{\mathrm{loc}}(\Sigma)$, *where* H *denotes the mean curvature of* Σ. *In the particular case where* Σ *has constant mean curvature*, *this formula reduces to*

$$D\tilde{H}_{|w=0}(v) = DH_{|w=0}((\tilde{N} \cdot N)v).$$

Proof. The implicit function theorem can be applied to the equation

$$p + tN(p) = q + s\tilde{N}(q)$$

to express (at least locally) p and t as functions of q and s, namely,

$$p = \Phi(q, s) \quad \text{and} \quad t = \Psi(q, s),$$

with $\Phi(q, 0) = q$ and $\Psi(q, 0) = 0$. It is easy to check that

$$\partial_s \Phi(\,\cdot\,, 0) = \tilde{N}^T \quad \text{and} \quad \partial_s \Psi(\,\cdot\,, 0) = \tilde{N} \cdot N,$$

where superscript T denotes the projection over $T\Sigma$.

Differentiating the identity

$$H_{\Psi(\,\cdot\,,w)}(\Phi(q,w(q))) = \tilde{H}_w(q)$$

with respect to w, at $w=0$, we find

$$DH_{|w=0}(\partial_s\Psi(\,\cdot\,,0)v) + \nabla H_{|w=0}\cdot\partial_s\Phi v = D\tilde{H}_{|w=0}(v).$$

The result then follows from the expression of $\partial_s\Phi$ and $\partial_s\Psi$ and the fact that $\tilde{H}_{|w=0} = H_{|w=0}$. □

3. Harmonic extensions

For all $x \in \mathbb{R}^2$ and all $r > 0$ we denote by $D(x,r) \subset \mathbb{R}^2$ the open disc of radius r centered at x and by $\overline{D}(x,r) \subset \mathbb{R}^2$ the closed disc of radius r centered at x. In this section, we study the harmonic extension of a function which is defined on the unit circle S^1 to a half cylinder $[0,\infty)\times S^1$, or to a punctured disc $\overline{D}^*(0,1)$, or to the complement of the closed unit disc $\mathbb{R}^2 - D(0,1)$. We use the fact that all these domains are conformal to each other and that the Laplacian is conformally invariant in dimension 2.

Let us assume that we are given a function $f \in \mathscr{C}^{2,\alpha}(S^1)$. We consider F to be the bounded harmonic extension of f to the half cylinder, endowed with the cylindrical metric

$$g_{\mathrm{cyl}} = ds^2 + d\theta^2.$$

In other words, F is bounded and is a solution of

$$\Delta_{g_{\mathrm{cyl}}} F = 0$$

in $[0,\infty)\times S^1$ with $F = f$ on $\{0\}\times S^1$.

Observe that one can use cylindrical coordinates to parametrize the punctured unit disc by

$$\tilde{X}(s,\theta) = (e^{-s}\cos\theta, e^{-s}\sin\theta),$$

in which case the function $\tilde{F}$ defined by $\tilde{F}\circ\tilde{X} := F$ is the unique bounded solution of

$$\Delta\tilde{F} = 0$$

(where Δ denotes the Laplacian in $\mathbb{R}^2$) in the punctured unit disc with $\tilde{F} = f$ on S^1. We set

$$W_f^{\mathrm{ins}} := \tilde{F}.$$

Also, one can use cylindrical coordinates to parametrize the complement of the unit disc in $\mathbb{R}^2$ by

$$\hat{X}(s,\theta) = (e^s\cos\theta, e^s\sin\theta),$$

in which case $\hat F$, defined by $\hat F \circ \hat X = F$, is the unique bounded solution of

$$\Delta \hat F = 0$$

in the complement of the unit disc with $\hat F = f$ on S^1. We set

$$W_f^{\mathrm{out}} := \hat F.$$

All properties of F will transfer easily to $\tilde F$ and $\hat F$.

Given a function f defined on S^1, we shall frequently assume that one or both of the following assumptions are fulfilled:

(H1) $$\int_{S^1} f\, d\theta = 0.$$

(H2) $$\int_{S^1} \cos\theta f\, d\theta = \int_{S^1} \sin\theta f\, d\theta = 0.$$

The following result follows essentially from [Fakhi and Pacard 2000], where a similar result was proven in higher dimensions.

Lemma 3.1. *There exists a constant $C > 0$ such that, for all $f \in \mathscr{C}^{2,\alpha}(S^1)$ satisfying (H1), we have*

$$\|e^s F\|_{\mathscr{C}^{2,\alpha}([0,\infty)\times S^1)} \le C \|f\|_{\mathscr{C}^{2,\alpha}(S^1)},$$

and, if f satisfies (H1) and (H2), we have

$$\|e^{2s} F\|_{\mathscr{C}^{2,\alpha}([0,\infty)\times S^1)} \le C \|f\|_{\mathscr{C}^{2,\alpha}(S^1)}.$$

Before we proceed with the proof of this result, let us emphasize that the norms in $\mathscr{C}^{2,\alpha}([0,\infty)\times S^1)$ are computed with respect to the cylindrical metric g_{cyl}.

Proof. We consider the Fourier series decomposition of the function f

$$f(\theta) = \sum_{n\in\mathbb{Z}} f_n e^{in\theta}.$$

Observe that $f_0 = 0$ when (H1) is fulfilled and $f_{\pm 1} = 0$ when (H2) is fulfilled. For the time being, let us assume that both (H1) and (H2) are satisfied. Then the (bounded) harmonic extension of f is given explicitly by

$$F(s,\theta) = \sum_{|n|\ge 2} e^{-|n|s} f_n e^{in\theta}.$$

Since

$$|f_n| \le \|f\|_{L^\infty(S^1)},$$

we get the pointwise estimate

$$|F(s,\theta)| \le 2\|f\|_{L^\infty(S^1)} \sum_{n\ge 2} e^{-ns} \le 2\|f\|_{L^\infty(S^1)} \frac{e^{-2s}}{1-e^{-s}},$$

which implies that

$$\sup_{[1,\infty)\times S^1} e^{2s}|F(s,\theta)| \leq C\|f\|_{L^\infty(S^1)}.$$

Increasing the value of $C > 0$ if necessary, we can use the maximum principle in the annular region $[0, 1] \times S^1$ to get

$$\sup_{[0,\infty)\times S^1} e^{2s}|F(s,\theta)| \leq C\|f\|_{L^\infty(S^1)}.$$

The estimates for the derivatives of F then follow from classical elliptic estimates, since Schauder's estimates can be applied on each annulus $[s, s+1] \times S^1$ for all $s \geq 0$. This already completes the proof of the result when both (H1) and (H2) are fulfilled. When only (H1) holds, one has to take into account the function $f_{\pm 1} e^{-s} e^{\pm i\theta}$, which accounts for the slower decay of F as e^{-s}. □

4. The Delaunay nodoids

Parametrization and notations. The Delaunay nodoid $\mathfrak{D}_\tau$ is a surface of revolution which can be parametrized by

$$X_\tau(s,\theta) := (\phi_\tau(s)\cos\theta, \phi_\tau(s)\sin\theta, \psi_\tau(s)), \tag{4-1}$$

where $(s,\theta) \in \mathbb{R} \times S^1$. Here the functions ϕ_τ and ψ_τ depend on the real parameter $\tau > 0$, but, unless necessary, we do not make this apparent in the notation. The function ϕ is chosen to be the unique, smooth, periodic, nonconstant solution of

$$\dot\phi^2 + (\phi^2 - \tau)^2 = \phi^2, \tag{4-2}$$

which takes its minimum value at $s = 0$ (we denote by $\cdot$ differentiation with respect to the parameter s), and the function ψ is obtained by integration of

$$\dot\psi = \phi^2 - \tau \tag{4-3}$$

with initial condition $\psi(0) = 0$. Observe that ϕ is a smooth solution of

$$\ddot\phi + 2\phi(\phi^2 - \tau) = \phi. \tag{4-4}$$

Since $\phi^2 - \tau$ changes sign, the function ψ is not monotone, and closer inspection of the solutions shows that $\mathfrak{D}_\tau$ is actually not embedded. The Delaunay nodoids also arise as the surface of revolution whose generating curve is a *roulette* of a hyperbola; we refer to [Eells 1987] for a description of this construction. The quantity $(1/4)\tau$ is sometimes referred to as the vertical flux of the Delaunay surface $\mathfrak{D}_\tau$; see [Rossman 2005, Definition 3.1]. Define

$$\underline{\tau} := \frac{\sqrt{1+4\tau}-1}{2} \quad \text{and} \quad \overline{\tau} := \frac{\sqrt{1+4\tau}+1}{2}, \tag{4-5}$$

which, thanks to (4-2) and (4-4), are, respectively, the minimum and maximum values of ϕ. As already mentioned, the function ϕ is periodic. We agree that s_τ denotes one *half* of the fundamental period of ϕ. Using (4-2), we can write

$$s_\tau = \int_{\underline{\tau}}^{\bar{\tau}} \frac{d\zeta}{\sqrt{\zeta^2 - (\zeta^2 - \tau)^2}}. \tag{4-6}$$

In the above parametrization, the induced metric on $\mathfrak{D}_\tau$ is given by

$$g_\tau := \phi^2(ds^2 + d\theta^2),$$

and it is easy to check that the second fundamental form on $\mathfrak{D}_\tau$ is given by

$$h_\tau := (\phi^2 + \tau)\, ds^2 + (\phi^2 - \tau) d\theta^2,$$

when the unit normal vector field is chosen to be

$$N_\tau := \frac{1}{\phi}((\tau - \phi^2)\cos\theta, (\tau - \phi^2)\sin\theta, \dot{\phi}).$$

Finally, the tensor k_τ is given by

$$k_\tau := \left(\phi + \frac{\tau}{\phi}\right)^2 ds^2 + \left(\phi - \frac{\tau}{\phi}\right)^2 d\theta^2.$$

In particular, the formulae for the induced metric and the second fundamental form imply that the mean curvature of this surface is constant and equal to

$$H := \tfrac{1}{2}\operatorname{tr}^g h = 1.$$

In these coordinates, it follows at once from the expression of g_τ and k_τ that the Jacobi operator about $\mathfrak{D}_\tau$ is given by

$$J_\tau := \frac{1}{2\phi^2}\left(\partial_s^2 + \partial_\theta^2 + 2\left(\phi^2 + \frac{\tau^2}{\phi^2}\right)\right).$$

Structure and refined asymptotics. The structure of the Delaunay surfaces $\mathfrak{D}_\tau$ is well understood, and it is known that, as the parameter τ tends to 0, $\mathfrak{D}_\tau$ converges to the union of infinitely many spheres of radius 1 which are arranged periodically along the vertical axis. To get a better grasp on the structure of $\mathfrak{D}_\tau$ as τ tends to 0, we have the following results, which were already used in many constructions of constant mean curvature surfaces by gluing [Mazzeo and Pacard 2001; Mazzeo et al. 2001; 2005]. For completeness we give independent proofs of these results.

Lemma 4.1. *As τ tends to* 0, *the following holds.*

(i) *The sequence of functions $\phi_\tau(\cdot + s_\tau)$ converges uniformly on compact sets of $\mathbb{R}$ to the function $s \mapsto (\cosh s)^{-1}$.*

(ii) *The sequence of functions* $\psi_\tau(\,\cdot+s_\tau)-\psi_\tau(s_\tau)$ *converges uniformly on compact sets of* $\mathbb{R}$ *to the function* $s\mapsto \tanh s$.

Proof. It is easy to check that $\phi_\tau(\,\cdot+s_\tau)$ is even and that $\phi_\tau(s_\tau)=\bar\tau$ converges to 1 as τ tends to 0 (this follows from the fact that the function ϕ_τ achieves its maximum value when $s=s_\tau$). Passing to the limit in (4-2), we conclude that the sequence of functions ϕ_τ converges uniformly on compact sets of $\mathbb{R}$ to a function ϕ_0 which is a solution of

$$\ddot\phi_0+2\phi_0^3=\phi_0.$$

Moreover, the function ϕ_0 is even and is equal to 1 when $s=0$. Therefore, necessarily $\phi_0(s)=(\cosh s)^{-1}$. Next, one can pass to the limit in (4-3) to prove that the sequence $\psi_\tau(\,\cdot+s_\tau)-\psi_\tau(s_\tau)$ converges to a function ψ_0 which is a solution of

$$\dot\psi_0=\phi_0^2$$

and satisfies $\psi_0(0)=0$. We find that $\psi_0(s)=\tanh s$. □

Now we investigate the behavior of $\mathfrak{D}_\tau$ close to the origin in $\mathbb{R}^3$. It turns out that the sequence of rescaled surfaces $(1/\tau)\mathfrak{D}_\tau$ converges on compact sets of $\mathbb{R}^3$ to a catenoid whose axis is the vertical axis:

Lemma 4.2. *As* τ *tends to* 0, *the following holds*:

(i) *The sequence of functions* $(1/\tau)\phi_\tau$ *converges uniformly on compact sets of* $\mathbb{R}$ *to the function* $s\mapsto\cosh s$.

(ii) *The sequence of functions* $(1/\tau)\psi_\tau$ *converges uniformly on compact sets of* $\mathbb{R}$ *to the function* $s\mapsto -s$.

Proof. It is easy to check that ϕ_τ is even and that $(1/\tau)\phi_\tau(0)=\underline{\tau}/\tau$ converges to 1 as τ tends to 0 (this follows from the fact that the function ϕ_τ achieves its minimum value when $s=0$). Passing to the limit in (4-2), we conclude that $(1/\tau)\phi_\tau$ converges uniformly on compact sets of $\mathbb{R}$ to a function ϕ_0 which is a solution of

$$\ddot\phi_0=\phi_0.$$

Moreover, ϕ_0 is even and is equal to 1 when $s=0$. Therefore, $\phi_0(s)=\cosh s$. Next, one can pass to the limit in (4-3) to prove that the sequence $(1/\tau)\psi_\tau$ converges to a function ψ_0 that is a solution of $\dot\psi_0=-1$ and satisfies $\psi_0(0)=0$. Therefore, we conclude that $\psi_0(s)=-s$, as desired. □

Geometrically, these results show that, as τ tends to 0, the Delaunay surface $\mathfrak{D}_\tau$ is close to infinitely many spheres of radius 1 which are arranged along the vertical axis (and are slightly overlapping), each sphere connected to its two neighbors by small rescaled catenoids.

We will need a refined and more quantitative version of Lemma 4.2. Observe that $\dot{\phi} < 0$ in $(-s_\tau, 0)$, and hence ϕ is a diffeomorphism from $(-s_\tau, 0)$ into $(\underline{\tau}, \bar{\tau})$. We can define the change of variables

$$r = \phi_\tau(s),$$

to express $s \in (-s_\tau, 0)$ as a function of $r \in (\underline{\tau}, \bar{\tau})$, so that the equality

$$X_\tau(s, \theta) = (r \cos\theta, r \sin\theta, u_\tau(r)),$$

where $r = \phi_\tau(s)$, holds for some function u_τ defined in an annulus of $\mathbb{R}^2$. Geometrically, this means that the image of $(-s_\tau, 0) \times S^1$ by X_τ is a vertical graph for some function u_τ which is defined over the annulus

$$\{x \in \mathbb{R}^2 : \underline{\tau} < |x| < \bar{\tau}\}.$$

Proposition 4.3. *As τ tends to* 0,

$$u_\tau(r) = \frac{\tau}{\sqrt{1+2\tau}} \log \frac{2r}{\tau} + \mathbb{O}_{\mathscr{C}^\infty}\left(\frac{\tau^3}{r^2}\right) + \mathbb{O}_{\mathscr{C}^\infty}(r^2)$$

for $r \in \left(2\underline{\tau}, \frac{1}{2}\bar{\tau}\right)$ uniformly as τ tends to 0.

The notation $f_1 = \mathbb{O}_{\mathscr{C}^\infty}(f_2)$ means that the function f_1 and all its derivatives with respect to the vector fields $r\partial_r$ and ∂_θ are bounded by a constant (depending on the order of derivation) times the (positive) function f_2.

Proof. By definition $\underline{\tau}$ is the minimum value of ϕ. Hence we can write

$$\phi(s) = \underline{\tau} \cosh(w(s))$$

for some function w which vanishes at $s = 0$. Plugging this into (4-2), we find that the function w is a solution of

$$\dot{w}^2 = 1 + 2\tau - \underline{\tau}^2(1 + \cosh^2 w).$$

As long as $|w(s) - s| \le 1$, we can estimate

$$w(s) = \sqrt{1+2\tau}\, s + \mathbb{O}(\tau^2 \cosh^2 s).$$

In particular, we conclude a posteriori that $|w(s) - s| \le 1$ holds, and hence that the above estimate is justified, provided $|s| \le -\log\tau - c$ for some constant $c > 0$ independent of $\tau \in (0, 1)$. In the range of study, we are entitled to consider the change of variable

$$r = \underline{\tau} \cosh w(s),$$

and express $s < 0$ as a function of r. We find

$$\sqrt{1+2\tau}\, s = -\log \frac{2r}{\underline{\tau}} + \mathbb{O}\left(\frac{\tau^2}{r^2}\right) + \mathbb{O}(r^2). \tag{4-7}$$

Finally, using (4-3), we can write

$$\dot{\psi} = -\tau + \underline{\tau}^2 \cosh^2 w.$$

Integrating over s, we get

$$\psi(s) = -\tau s + \mathbb{O}(\tau^2 \cosh^2 s),$$

and the result follows directly from (4-7) together with the fact that $u_\tau(r) = \psi(s)$, by definition, Similar estimates can be obtained for the derivatives of u_τ. □

A close inspection of the proof of Proposition 4.3 also yields the following.

Lemma 4.4. *As τ tends to* 0, *half of the fundamental period of the function ϕ_τ can be expanded as*

$$s_\tau = -\log \tau + \mathbb{O}(1),$$

and there exists a constant $C > 1$ such that

$$\frac{\tau}{C} \cosh s \le \phi_\tau \le C\tau \cosh s,$$

when $s \in (-s_\tau, s_\tau)$, this estimate being uniform as τ tends to 0.

Proof. The asymptotics of the half period of ϕ can also be derived from the formula (4-6). The estimate for ϕ follows from the proof of Proposition 4.3. □

Analysis of the Jacobi operator. We analyze the mapping properties of the Jacobi operator about the Delaunay surface $\mathfrak{D}_\tau$, paying special attention to what happens when τ tends to 0. This analysis is very close to the one available in [Mazzeo and Pacard 2001; Hauswirth and Pacard 2007]. Again, we give a self-contained proof adapted to the nonlinear argument we use in subsequent sections.

We first analyze the behavior, as τ tends to 0, of the potential which appears in the expression of J_τ. To this end, we assume that, for each $\tau > 0$, we are given a real number $t_\tau \in \mathbb{R}$ and we define

$$\xi_\tau := \left(\phi_\tau^2 + \frac{\tau^2}{\phi_\tau^2}\right)(\cdot - t_\tau).$$

Lemma 4.5. *As τ tends to* 0, *a subsequence of the sequence of functions ξ_τ converges uniformly on compact sets of $\mathbb{R}$ either to* 0 *or to the function $s \mapsto (\cosh(s - s_0))^{-2}$, for some $s_0 \in \mathbb{R}$.*

Proof. We define

$$\zeta_\tau := \left(\phi_\tau + \frac{\tau}{\phi_\tau}\right)(\cdot - t_\tau).$$

Using (4-2), we find that ζ_τ is a solution of

$$\dot{\zeta}_\tau^2 = (\zeta_\tau^2 - 2\tau)(1 + 4\tau - \zeta_\tau^2), \tag{4-8}$$

and, additionally, (4-4) implies that

$$\ddot{\zeta}_\tau = \zeta_\tau(1+6\tau-2\zeta_\tau^2). \tag{4-9}$$

Now we can estimate

$$\zeta_\tau^2 = \left(\phi - \frac{\tau}{\phi}\right)^2 + 4\tau \le 1+4\tau,$$

where we have used (4-2), which provides the estimate $(\phi^2-\tau)^2 \le \phi^2$. This implies that ζ_τ and its derivatives remain bounded as τ tends to 0. We can then let τ tend to 0 and pass to the limit in (4-8) and (4-9) to get that, as τ tends to 0, the sequence ζ_τ converges on compact sets to a solution of the equation $\ddot{\zeta} = \zeta(1-2\zeta^2)$, which satisfies $\dot{\zeta}^2 = \zeta^2(1-\zeta^2)$. Hence ζ is either 0 or a translation of $z \mapsto (\cosh s)^{-1}$. The result then follows from the identity $\xi_\tau = \zeta_\tau^2 - 2\tau$. □

We denote by $\pm\delta_j(\tau)$, for $j \in \mathbb{N}$, the indicial roots of the operator J_τ. Recall that the indicial roots $\pm\delta_j$ correspond to the indicial roots of

$$J_{\tau,j} := \frac{1}{2\phi^2}\left(\partial_t^2 - j^2 + 2\left(\phi^2 + \frac{\tau^2}{\phi^2}\right)\right),$$

which appears in the Fourier decomposition of the operator J_τ in the θ variable. The indicial roots of $J_{\tau,j}$ characterize the exponential growth or decay rate at infinity of the solutions of the homogeneous problem $J_{\tau,j} w = 0$. In general, it is a very hard problem to determine the exact value of the indicial roots of an operator, but in the present case, taking advantage of the geometric nature of the problem, we can prove the following.

Proposition 4.6. *For all $\tau > 0$, we have $\delta_0(\tau) = \delta_1(\tau) = 0$. Furthermore, for $j \ge 2$,*

$$\delta_j(\tau) \ge \sqrt{j^2 - 2 - 4\tau},$$

provided $\tau < \sqrt{j^2-2}$.

Proof. The fact that $\delta_0(\tau) = 0$ follows from the observation that the function $\dot{\phi}/\phi$ is periodic and solves

$$J_{\tau,0}(\dot{\phi}/\phi) = 0.$$

This follows from direct computation, or it can be derived from the fact that

$$\Phi_0^+ := \dot{\phi}/\phi$$

is the Jacobi field associated to vertical translation (see page 134). Since the function ϕ is periodic, the homogeneous problem $J_{\tau,0} w = 0$ has a bounded solution, which implies that $\delta_0(\tau) = 0$.

The fact that $\delta_1(\tau) = 0$ follows from the observation that the function $\phi - \tau/\phi$ is periodic and solves

$$J_{\tau,1}\Big(\phi - \frac{\tau}{\phi}\Big) = 0.$$

Again, this follows from direct computation or can be derived from the fact that

$$\Phi_1^+ := \Big(\phi - \frac{\tau}{\phi}\Big)\cos\theta \quad \text{and} \quad \Phi_{-1}^+ := \Big(\phi - \frac{\tau}{\phi}\Big)\sin\theta$$

are the Jacobi fields associated to translations perpendicular to the axis of the Delaunay surface.

The estimate from below for the other indicial roots follows from the fact that, according to (4-2),

$$2\Big(\phi^2 + \frac{\tau^2}{\phi^2}\Big) = 2\Big(\phi - \frac{\tau}{\phi}\Big)^2 + 4\tau \le 2 + 4\tau.$$

This, in particular, implies that the potential in $\partial_s^2 - j^2 + 2(\phi^2 + \tau^2/\phi^2)$ can be estimated from below by $\bar{\delta}_j^2$, where

$$\bar{\delta}_j := \sqrt{j^2 - 2 - 4\tau}.$$

The result then follows from the maximum principle and standard ODE arguments, since the function $s \mapsto e^{\bar{\delta}_j s}$ can be used as a barrier to prove the existence of two positive solutions of $J_{\tau,j} w = 0$ which are defined on $(0, \infty)$, one being bounded from above by $e^{-\bar{\delta}_j s}$ and the other from below by $e^{\bar{\delta}_j s}$. This implies that $\delta_j \ge \bar{\delta}_j$, completing the proof of the result. □

For all $\delta \in \mathbb{R}$, we define the operator

$$\begin{aligned} L_\delta : e^{\delta s}\mathscr{C}^{2,\alpha}(\mathbb{R} \times S^1) &\to e^{\delta s}\mathscr{C}^{0,\alpha}(\mathbb{R} \times S^1), \\ w &\mapsto \phi^2 J_\tau w, \end{aligned}$$

where the norms in the function spaces $\mathscr{C}^{k,\alpha}(\mathbb{R} \times S^1)$ are computed with respect to the cylindrical metric g_{cyl}. Observe that the Jacobi operator has been multiplied by the conformal factor ϕ^2, and hence

$$\phi^2 J_\tau = \tfrac{1}{2}\Big(\partial_s^2 + \partial_\theta^2 + 2\Big(\phi^2 + \frac{\tau^2}{\phi^2}\Big)\Big).$$

Also, this operator depends on the parameter τ. We now study the mapping properties of $\phi^2 J_\tau$ as the parameter τ tends to 0. The following result selects a range of weights for which the norm of the solution of $L_\delta w = f$ is controlled, uniformly as τ tends to 0.

Proposition 4.7. *Assume that $|\delta| > 1$, $\delta \notin \mathbb{Z}$ is fixed. Then there exist $\tau_\delta > 0$ and $C > 0$, only depending on δ, such that, for all $\tau \in (0, \tau_\delta)$ and all $w \in e^{\delta s}\mathscr{C}^{2,\alpha}(\mathbb{R} \times S^1)$, we have*

$$\|e^{-\delta s} w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \leq C \|e^{-\delta s} L_\delta w\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}.$$

Proof. Observe that, thanks to Schauder's elliptic estimates, it is enough to prove

$$\|e^{-\delta s} w\|_{L^\infty(\mathbb{R}\times S^1)} \leq C \|e^{-\delta s} \phi_\tau^2 J_\tau w\|_{L^\infty(\mathbb{R}\times S^1)},$$

provided τ is close enough to 0. The proof of this estimate is by contradiction. Assume that, for some sequence τ_n tending to 0, there exists a sequence of functions w_n such that

$$\|e^{-\delta s} w_n\|_{L^\infty(\mathbb{R}\times S^1)} = 1 \quad \text{and} \quad \lim_{j\to\infty} \|e^{-\delta s} \phi_{\tau_n}^2 J_{\tau_n} w_n\|_{L^\infty(\mathbb{R}\times S^1)} = 0.$$

Pick a point $s_n \in \mathbb{R}$ such that $\|e^{-\delta s_n} w_n(s_n, \cdot)\|_{L^\infty(S^1)} \geq \frac{1}{2}$ and define the rescaled sequence

$$\bar{w}_n(s, \theta) := e^{-\delta s_n} w_n(s + s_n, \theta).$$

We still have

$$\|e^{-\delta s} \bar{w}_n\|_{L^\infty(\mathbb{R}\times S^1)} = 1 \quad \text{and} \quad \lim_{j\to\infty} \|e^{-\delta s} \bar{L}_n \bar{w}_n\|_{L^\infty(\mathbb{R}\times S^1)} = 0,$$

where, by definition, $\bar{L}_n$ is defined by

$$\bar{L}_n := \partial_s^2 + \partial_\theta^2 + 2\left(\phi_{\tau_n}^2 + \frac{\tau_n^2}{\phi_{\tau_n}^2}\right)(\,\cdot + s_n).$$

Elliptic estimates and the Ascoli–Arzelà theorem allow us to extract some subsequence and pass to the limit as n tends to ∞ to get a function w_∞ which is a nontrivial solution to either

$$(\partial_s^2 + \partial_\theta^2) w_\infty = 0 \tag{4-10}$$

or

$$\left(\partial_s^2 + \partial_\theta^2 + \frac{2}{\cosh^2(\,\cdot + s_*)}\right) w_\infty = 0, \tag{4-11}$$

according to the different cases described in Lemma 4.5. To simplify notation, we assume that $s_* = 0$; straightforward modifications are needed to handle the general case. Observe that we also have

$$\|e^{-\delta s} w_\infty\|_{L^\infty(\mathbb{R}\times S^1)} \leq 1,$$

and $\|w_\infty(0, \cdot)\|_{L^\infty(S^1)} \geq \frac{1}{2}$.

We decompose w_∞ as

$$w_\infty(s,\theta)=\sum_{j\in\mathbb{Z}} w^{(j)}(s)e^{ij\theta}.$$

It is easy to prove that, for any solution of (4-10), $w^{(j)}$ is a linear combination of $e^{\pm js}$ and it is not bounded by a constant times $e^{\delta s}$ unless δ is an integer (which we have assumed not to be the case).

Similarly, if w_∞ is a solution of (4-11), we find that $w^{(j)}$ is a solution of

$$\left(\partial_s^2-j^2+\frac{2}{\cosh^2 s}\right)w^{(j)}=0, \tag{4-12}$$

and is asymptotic to either e^{js} or to e^{-js} at $\pm\infty$. Inspection of the behavior of (4-12) at infinity then implies that there is no solution bounded by a constant times $e^{\delta s}$ if $|j|<|\delta|$. When $|j|>|\delta|$, inspection of the behavior of (4-12) at infinity implies that any solution is necessarily bounded, and the maximum principle then implies that this solution is identically 0 (observe that in this case $j^2>2$ since $|j|>|\delta|>1$, and hence the potential in (4-12) is negative).

When $j=0$, all solutions of

$$\left(\partial_s^2+\frac{2}{\cosh^2 s}\right)w^{(0)}=0$$

are linear combinations of the functions $\tanh s$ and $1-s\tanh s$ and none is bounded by a constant times $e^{\delta s}$ unless $\delta=0$ (which is not the case).

Finally — and this is the reason we had to choose $|\delta|>1$ — when $j=1$, all solutions of

$$\left(\partial_s^2-1+\frac{2}{\cosh^2 s}\right)w^{(1)}=0$$

are linear combinations of the functions $(\cosh s)^{-1}$ and $s(\cosh s)^{-1}+\sinh s$ and none is bounded by a constant times $e^{\delta s}$ unless $|\delta|\le 1$ (which is contrary to our assumption). Again we have reached a contradiction. Having reached a contradiction in all cases, the proof of the Proposition is complete. □

Thanks to the previous result, we can now describe the mapping properties of $\phi^2 J_\tau$ for the range of weights δ of interest for our problem.

Proposition 4.8. *Assume that $|\delta|>1$, $\delta\notin\mathbb{Z}$ is fixed. Then there exist $\tau_\delta>0$ and $C>0$, only depending on δ, such that, for all $\tau\in(0,\tau_\delta)$, the operator L_δ is an isomorphism, the norm of whose inverse is bounded independently of τ.*

Proof. Injectivity follows at once from Proposition 4.7. As far as surjectivity is concerned, we give here a simple self-contained proof in the case where $\delta\in(1,\sqrt{2})$ (or $\delta\in(-\sqrt{2},-1)$). We then sketch a general proof.

To fix the ideas, assume that $\delta\in(1,\sqrt{2})$. First assume that $f\in\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)$

has compact sets support and decompose it as

$$f(s,\theta) = f_0(s) + f_{\pm 1}(s)e^{\pm i\theta} + \bar{f}(s,\theta),$$

where, by definition,

$$\bar{f} := \sum_{j \neq 0, \pm 1} f_j(s) e^{ij\theta}.$$

Observe that, if we restrict our attention to functions $\bar{w}$ whose Fourier decomposition in the θ variable is of the form

$$\bar{w}(s,\theta) = \sum_{j \neq 0, \pm 1} w_j(s) e^{ij\theta},$$

we have

$$\int_{\mathbb{R}\times S^1} \left(|\partial_s \bar{w}|^2 + |\partial_\theta \bar{w}|^2 - 2\left(\phi^2 + \frac{\tau^2}{\phi^2}\right)\bar{w}^2 \right) ds\, d\theta \geq (2 - 4\tau) \int_{\mathbb{R}\times S^1} \bar{w}^2\, ds\, d\theta. \tag{4-13}$$

This follows at once from the estimate of the potential involved in the expression of J_τ obtained in the proof of Proposition 4.6, namely,

$$2\left(\phi^2 + \frac{\tau^2}{\phi^2}\right) \leq 2 + 4\tau \tag{4-14}$$

together with the fact that

$$\int_{\mathbb{R}\times S^1} |\partial_\theta \bar{w}|^2\, ds\, d\theta \geq 4 \int_{\mathbb{R}\times S^1} \bar{w}^2\, ds\, d\theta.$$

Thus, if we assume that $\sqrt{2}\tau < 1$, this inequality implies that we can solve

$$\phi^2 J_\tau \bar{w} = \bar{f}$$

in $H^1(\mathbb{R}\times S^1)$. Elliptic estimates then imply that $\bar{w} \in \mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)$. Finally, the solvability of

$$\phi^2 J_\tau (w_j e^{ij\theta}) = f_j e^{ij\theta},$$

for $j = 0, \pm 1$, follows easily from integration of the associated second order ordinary differential equation starting from $-\infty$. Hence $w_j \equiv 0$ when s is close to $-\infty$. Obviously, the function

$$w := w_0 + w_{\pm 1} e^{\pm i\theta} + \bar{w}$$

is a solution of the equation $\phi^2 J_\tau w = f$.

We claim that, provided τ is chosen small enough, $w \in e^{\delta s}\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)$. Assuming that the claim is already proven, Proposition 4.7 applies and we get

$$\|e^{-\delta s} w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \leq C \|e^{-\delta s} f\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}$$

for any function f having compact support. The general result, when f does not necessarily have compact support, follows from a standard exhaustion argument. We choose a sequence of functions $f^{(n)} \in \mathscr{C}^{0,\alpha}(\mathbb{R} \times S^1)$ having compact support converging on compact sets to a given function $f \in e^{\delta s}\mathscr{C}^{0,\alpha}(\mathbb{R} \times S^1)$. Moreover, without loss of generality, we can assume that

$$\|e^{-\delta s} f^{(n)}\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)} \leq C \|e^{-\delta s} f\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}$$

for some constant $C > 0$ independent of $n \geq 0$. Thanks to the above, we have a sequence of solutions of $\phi^2 J_\tau w^{(n)} = f^{(n)}$ satisfying

$$\|e^{-\delta s} w^{(n)}\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \leq C \|e^{-\delta s} f\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}.$$

Extracting some subsequence and passing to the limit, one gets the existence of $w \in e^{\delta s}\mathscr{C}^{0,\alpha}(\mathbb{R} \times S^1)$, a solution of $\phi^2 J_\tau w = f$ satisfying

$$\|e^{-\delta s} w\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)} \leq C \|e^{-\delta s} f\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}.$$

The result then follows from Schauder's estimates.

It remains to prove the claim. We keep the notations introduced above. We first prove that $\bar{w}$ tends to 0 exponentially fast at infinity. Indeed, away from the support of $\bar{f}$, we can multiply the equation $\phi^2 J_\tau \bar{w} = \bar{f}$ by $\bar{w}$ and integrate over S^1 to get

$$\frac{1}{2}\frac{d^2}{ds^2}\left(\int_{S^1} \bar{w}^2 \, d\theta\right) = \int_{S^1} \left(|\partial_s \bar{w}|^2 + |\partial_\theta \bar{w}|^2 - \left(\phi^2 + \frac{\tau^2}{\phi^2}\right)\bar{w}^2 \right) d\theta.$$

But

$$\int_{S^1} |\partial_\theta \bar{w}|^2 \, d\theta \geq 2 \int_{S^1} \bar{w}^2 \, d\theta,$$

and we conclude from (4-14) that

$$\frac{d^2}{ds^2}\left(\int_{S^1} \bar{w}^2 \, d\theta\right) \geq 4(1-2\tau) \int_{S^1} \bar{w}^2 \, d\theta.$$

Since we have assumed that $\delta \in (1, \sqrt{2})$, we can assume that $\tau > 0$ is small enough so that

$$\delta^2 \leq 2(1-2\tau),$$

and using the fact that $\bar{w}$ is bounded, we conclude that there exists $C > 0$ such that

$$\int_{S^1} \bar{w}^2 \, d\theta \leq C (\cosh s)^{-2\delta}.$$

This shows that $\bar{w} \in (\cosh s)^{-\delta} L^2(\mathbb{R} \times S^1)$, and, by elliptic regularity, this implies that $\bar{w} \in (\cosh s)^{-\delta}\mathscr{C}^{2,\alpha}(\mathbb{R} \times S^1)$.

It remains to check that the functions w_0 and $w_{\pm 1}$ are at most growing linearly at $+\infty$. This follows at once from the fact that, for s large enough, these functions are solutions of the second order homogeneous ordinary differential equations

$$\left(\partial_s^2 - j^2 + 2\left(\phi^2 + \frac{\tau^2}{\phi^2}\right)\right) w_j = 0.$$

For $j = 0$, 1, this ordinary differential equation, whose potential is periodic, has one solution which is periodic (see the proof of Proposition 4.6), and a standard result implies that the other linearly independent solution of this ordinary differential equation is at most growing linearly (see the appendix). In particular, $w_j \in e^{\delta s}\mathscr{C}^{2,\alpha}(\mathbb{R})$, and this completes the proof of the claim.

We briefly explain how the proof of the general result can be obtained. The idea is to solve the equation $\phi^2 J_\tau \bar{w}_{s_0} = \bar{f}$ in $[-s_0, s_0] \times S^1$ with 0 boundary conditions. This can be done using the coercivity inequality (4-13). Then the proof of Proposition 4.7 can be adapted to prove that

$$\|e^{-\delta s}\bar{w}_{s_0}\|_{\mathscr{C}^{2,\alpha}([-s_0,s_0]\times S^1)} \le C\|e^{-\delta s}\bar{f}\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}$$

for some constant $C > 0$ independent of $s_0 > 1$ (observe that we use the fact that the Fourier decomposition of the function $\bar{w}$ in the θ variable does not have any component over 1 and $e^{\pm i\theta}$). It then remains to pass to the limit in the sequence w_{s_0} as s_0 tends to ∞ to prove the existence of a solution $\bar{w}$ to $\phi^2 J_\tau \bar{w} = \bar{f}$ in $\mathbb{R} \times S^1$, which satisfies the correct estimate. □

Using similar arguments, one can give a direct proof of the following general result, which will not be needed in this paper.

Theorem 4.9. *Assume that $\delta \neq \pm\delta_j(\tau)$ for all $j \in \mathbb{N}$. Then L_δ is an isomorphism.*

The proof of this result follows from the general theory developed in [Pacard 2008] (see Theorem 10.2.1 on page 61 and Proposition 12.2.1 on page 81) or [Melrose 1993; Mazzeo 1991].

In what follows we restrict our attention to functions which are invariant under some symmetries. More precisely, we assume that the functions are invariant under the action on S^1 of the dihedral group $\mathrm{Dih}^{(2)}_{m+1}$ of isometries of $\mathbb{R}^2$ which leave a regular polygon with $m+1$ sides fixed. The operator associated to $\phi^2 J_\tau$, acting on the weighted space of functions which are invariant under these symmetries, is denoted by $L^\sharp_\delta$. This time $L^\sharp_\delta$ is an isomorphism provided $\delta \neq \pm\delta_j$ for all $j \in \mathbb{Z}$ for which there exist eigenfunctions of ∂^2_θ which are invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$, namely, $j \notin m\mathbb{Z}$. Observe that, when $j = 1$, there are no such eigenfunctions, and hence working equivariantly allows us to extend the range in which the weight parameter δ can be chosen.

Close inspection of the previous proof shows that the range in which the weight δ can be chosen so that the inverse of $L^\sharp_\delta$ remains bounded as τ tends to 0 can be enlarged if we work equivariantly. Even though we will not use it, we state here the corresponding result for the sake of completeness.

Proposition 4.10. *Assume that $\delta \notin m\mathbb{Z}$ is fixed. Then there exist $\tau_\delta > 0$ and $C > 0$, only depending on δ, such that, for all $\tau \in (0, \tau_\delta)$ and for all $w \in e^{\delta s}\mathscr{C}^{2,\alpha}(\mathbb{R} \times S^1)$, which is invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$, we have*

$$\|e^{-\delta s} w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \leq C \|e^{-\delta s} L^\sharp_\delta w\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}.$$

The mean curvature of normal graphs over $\mathfrak{D}_\tau$. In this section, we investigate the mean curvature of a surface which is a normal graph over $\mathfrak{D}_\tau$. Given a smooth function w (small enough) defined on $\mathfrak{D}_\tau$, we consider the surface parametrized by

$$\tilde{X}(s, \theta) = X_\tau(s, \theta) + w(s, \theta) N_\tau(s, \theta).$$

We have the following technical result.

Lemma 4.11. *The mean curvature of the surface parametrized by $\tilde{X}$ is given by*

$$H(w) = 1 + J_\tau w + \frac{1}{\phi} Q_\tau\left(\frac{w}{\phi}\right),$$

where the second order differential nonlinear operator Q_τ depends on τ and satisfies

$$\|Q_\tau(v_2) - Q_\tau(v_1)\|_{\mathscr{C}^{0,\alpha}([s,s+1]\times S^1)} \\ \leq c\left(\|v_1\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)} + \|v_2\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)}\right)\|v_2 - v_1\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)}$$

for some constant $c > 0$ independent of s and $\tau \in (0, 1)$ and for all functions v_1, v_2 satisfying $\|v_i\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)} \leq 1$.

Proof. This follows at once from Proposition 2.1 together with the fact that the functions ϕ, τ/ϕ, and $\dot{\phi}/\phi$, as well as their derivatives, are uniformly bounded as τ tends to 0. Indeed, we have

$$g_w = g - 2wh + w^2 k = \phi^2\left(\left(1 - \left(\phi + \frac{\tau}{\phi}\right)\frac{w}{\phi}\right)^2 ds^2 + \left(1 - \left(\phi - \frac{\tau}{\phi}\right)\frac{w}{\phi}\right)^2 d\theta^2\right).$$

Hence $\phi^{-2} g_w$ has coefficients which are bounded functions of w/ϕ. Similarly, the tensor $\phi^{-1}(h - wk)$ also has coefficients which are bounded functions of w/ϕ. Using this, it is straightforward to check that the nonlinear terms in $H(w)$ are a function of $\partial_s^k \partial_\theta^l w/\phi$ for $k + l = 0, 1, 2$ with coefficients bounded by $1/\phi$. Finally, observe that

$$\frac{\partial_s w}{\phi} = \partial_s\left(\frac{w}{\phi}\right) + \frac{\dot{\phi}}{\phi}\frac{w}{\phi},$$

and hence any expressions of the form $\partial_s w/\phi$ can also be expressed as a linear combination (with coefficients bounded uniformly as τ tends to 0) of the function w/ϕ and its derivatives. We leave the details to the reader. □

A first fixed-point argument. We assume that we are given $\tau > 0$. We define $\bar{s} \in (-s_\tau, 0)$ by the identity

$$\phi_\tau(\bar{s}) = \tau^{3/4}.$$

Observe that $\bar{s}$ depends on τ even though we have chosen not to make this apparent in the notation. Moreover, it follows from the proof of Proposition 4.3 that

$$\bar{s} = \tfrac{1}{4} \log \tau + \mathcal{O}(1)$$

as τ tends to 0. We define the truncated nodoid $\mathfrak{D}_\tau^+$ to be the image of $[\bar{s}, +\infty) \times S^1$ by X_τ. Observe that this surface has a boundary, and, thanks to Proposition 4.3, close to this boundary it can be parametrized as the vertical graph of the function

$$x \mapsto \tau \log \frac{2|x|}{\tau} + \mathcal{O}_{\mathscr{C}^\infty}(\tau^{3/2})$$

over the annulus $\bar{D}(0, \tau^{3/4}) - D(0, \tau^{3/4}/2)$. Moreover, $\mathfrak{D}_\tau^+$ has one end in the upper half space.

In this section, we apply the implicit function theorem (or, to be more precise, a fixed-point argument for a contraction mapping) to produce an infinite-dimensional family of constant mean curvature surfaces which are close to $\mathfrak{D}_\tau^+$ and have one boundary which can be described using a function $f : S^1 \to \mathbb{R}$.

Proposition 4.12. *Assume that we are given $\kappa > 0$ large enough* (*to be fixed later on*). *For all $\tau > 0$ small enough and for all functions f invariant under the action of* $\mathrm{Dih}_{m+1}^{(2)}$ *satisfying* (H1) (*observe that* (H2) *is automatically satisfied*) *and*

$$\|f\|_{\mathscr{C}^{2,\alpha}(S^1)} \le \kappa \tau^{3/2}, \tag{4-15}$$

there exists a constant mean curvature surface $\mathfrak{D}_{\tau,f}^+$ with mean curvature equal to 1 *that is a graph over $\mathfrak{D}_\tau^+$ and has one Delaunay end asymptotic to the end of $\mathfrak{D}_\tau^+$ and one boundary. When $f = 0$, $\mathfrak{D}_{\tau,0} = \mathfrak{D}_\tau^+$ and, close to its boundary, the surface $\mathfrak{D}_{\tau,f}$ is a* vertical graph *over the annulus*

$$\{x \in \mathbb{R}^2 : \tfrac{1}{2}\tau^{3/4} \le |x| \le \tau^{3/4}\}$$

for the function $x \mapsto U_{\tau,f}^{\upharpoonleft}(\tau^{-3/4}x)$, which can be expanded as

$$U_{\tau,f}^{\upharpoonleft}(x) = \tau \log \frac{2}{\tau^{1/4}} + \tau \log |x| - W_f^{\mathrm{ins}}(x) + \overline{U}_{\tau,f}^{\upharpoonleft}(x), \tag{4-16}$$

where we recall that W_f^{ins} denotes the bounded harmonic extension of f in the punctured unit disc and where

$$\|\overline{U}_{\tau,0}^{\upharpoonleft}\|_{\mathscr{C}^{2,\alpha}(\bar{D}(0,1)-D(0,1/2))} \le C\tau^{3/2}. \tag{4-17}$$

Moreover, the nonlinear operator

$$\mathscr{C}^{2,\alpha}(S^1) \ni f \mapsto \overline{U}^{\upharpoonleft}_{\tau,f} \in \mathscr{C}^{2,\alpha}\big(\overline{D}(0,1) - D\big(0,\tfrac{1}{2}\big)\big)$$

is Lipschitz, and, given $\delta \in (-2,-1)$, *we have*

$$\|\overline{U}^{\upharpoonleft}_{\tau,f'} - \overline{U}^{\upharpoonleft}_{\tau,f}\|_{\mathscr{C}^{2,\alpha}(\overline{D}(0,1)-D(0,1/2))} \leq C\tau^{(2+\delta)/4}\|f'-f\|_{\mathscr{C}^{2,\alpha}(S^1)} \tag{4-18}$$

for some constant $C > 0$, *independent of* κ, τ *and* f, f'. *Finally,* $\mathfrak{D}^{+}_{\tau,f}$ *is invariant under the action of the dihedral group* $\mathrm{Dih}^{(2)}_{m+1}$.

Before we proceed with the proof, observe that we have chosen to describe the surface near its boundary as the graph of the function $x \mapsto U^{\upharpoonleft}_{\tau,f}(\tau^{-3/4}x)$, and, consequently, the function $U^{\upharpoonleft}_{\tau,f}$ is defined over the annulus $\overline{D}(0,1) - D\big(0,\tfrac{1}{2}\big)$. Alternatively, we could have chosen not to scale the coordinates and to have a function defined over the annulus $\overline{D}(0,\tau^{3/4}) - D(0,\tau^{3/4}/2)$, which would be more natural. However, with this latter choice, we would have to consider, in (4-17) and (4-18), function spaces where partial derivatives are taken with respect to the vector fields $r\partial_r$ and ∂_θ to evaluate the norm of these functions, while with the former choice, the Hölder spaces are the usual ones.

Proof. The proof of this result is fairly technical but, by now, standard. To begin with, in the annular region which is the image of $(\bar{s}-2,\bar{s}+2)\times S^1$ by X_τ, we modify the unit vector field N_τ into $\bar{N}_\tau$ in such a way that $\bar{N}_\tau$ is equal to $-e_3$, the downward pointing unit normal vector field on the image of $(\bar{s}-1,\bar{s}+1)\times S^1$ by X_τ. Using Proposition 2.2, direct estimates imply that the expression of the mean curvature given in Lemma 4.11 has to be altered to

$$\overline{H}(w) = 1 + J_\tau w + \frac{1}{\phi^2} l_\tau w + \frac{1}{\phi}\overline{Q}_\tau\left(\frac{w}{\phi}\right),$$

where $\overline{Q}_\tau$ enjoys properties which are similar to the properties enjoyed by Q_τ and where l_τ is a linear second order partial differential operator in ∂_s and ∂_θ whose coefficients are smooth, have support in $[\bar{s}-2,\bar{s}+2]\times S^1$, and are bounded (in the $\mathscr{C}^\infty$ topology) by a constant (independent of τ) times $\tau^{1/2}$. This estimate comes from the fact that

$$N\cdot(-e_3) = 1 + \mathbb{O}(\tau^{1/2})$$

on the image of $[\bar{s}-2,\bar{s}+2]\times S^1$ by X_τ.

We assume that we are given a function $f \in \mathscr{C}^{2,\alpha}(S^1)$ satisfying (H1), (H2), and (4-15), and we denote by F the harmonic extension of f in $(\bar{s},\infty)\times S^1$.

Given these data, we would like to solve the nonlinear equation

$$\phi^2 J_\tau(F+w) + l_\tau(F+w) + \phi\overline{Q}_\tau\left(\frac{F+w}{\phi}\right) = 0 \tag{4-19}$$

in $(\bar{s}, \infty) \times S^1$. Provided w is small enough and decays exponentially at infinity, this will then provide constant mean curvature surfaces which are close to a half nodoid $\mathfrak{D}_\tau^+$.

We choose an extension operator

$$\mathscr{E}_\tau : \mathscr{C}^{0,\alpha}([\bar{s}, \infty) \times S^1) \to \mathscr{C}^{0,\alpha}(\mathbb{R} \times S^1)$$

such that

$$\mathscr{E}_\tau(\psi) = \begin{cases} \psi & \text{in } [\bar{s}, \infty) \times S^1, \\ 0 & \text{in } (-\infty, \bar{s}-1] \times S^1, \end{cases}$$

and

$$\|\mathscr{E}_\tau(\psi)\|_{\mathscr{C}^{0,\alpha}([\bar{s}-1,\bar{s}+1]\times S^1)} \le C\|\psi\|_{\mathscr{C}^{0,\alpha}([\bar{s},\bar{s}+1]\times S^1)}.$$

We rewrite (4-19) as

$$\phi^2 J_\tau w = -\mathscr{E}_\tau\left(\phi^2 J_\tau(F+w) + l_\tau F + \phi\bar{Q}_\tau\left(\frac{F+w}{\phi}\right)\right), \tag{4-20}$$

where, this time, the function w is defined on all $\mathbb{R} \times S^1$ (to be more precise, one should say that, on the right side, we consider the restriction of w to $[\bar{s}, \infty) \times S^1$).

The following estimates follow easily if one uses the fact that

$$\frac{C}{\tau}\cosh s \le \phi \le C\tau \cosh s \quad \text{in } (-s_\tau, s_\tau)$$

for some $C > 1$, and also that ϕ is periodic of period $2s_\tau$. Assume that $\delta \in (-2, -1)$ is fixed. It is easy to check that there exists a constant $c > 0$ (independent of κ) and a constant $c_\kappa > 0$ (depending on κ) such that

$$\left\| e^{-\delta s}\mathscr{E}_\tau\left(\left(\phi^2 + \frac{\tau^2}{\phi^2}\right)F\right)\right\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)} \le c\tau^{1/2}\|f\|_{\mathscr{C}^{2,\alpha}(S^1)},$$
$$\|e^{-\delta s}\mathscr{E}_\tau(l_\tau(F+w))\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)} \le c\tau^{1/2}\big(\|e^{-\delta s}w\|_{\mathscr{C}^{2,\alpha}_\delta(\mathbb{R}\times S^1)} + \tau^{-\delta/4}\|f\|_{\mathscr{C}^{2,\alpha}(S^1)}\big),$$

and we also have

$$\left\| e^{-\delta s}\mathscr{E}_\tau\left(\phi Q_\tau\left(\frac{w'+F'}{\phi}\right) - \phi Q_\tau\left(\frac{w+F}{\phi}\right)\right)\right\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}$$
$$\le c_\kappa(\tau^{3/4}\|e^{-\delta s}(w'-w)\|_{\mathscr{C}^{2,\alpha}_\delta(\mathbb{R}\times S^1)} + \tau^{(3-\delta)/4}\|f'-f\|_{\mathscr{C}^{2,\alpha}(S^1)}),$$

provided w and w' satisfy

$$\|e^{-\delta s}w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} + \|e^{-\delta s}w'\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le C_\kappa\tau^2$$

for some fixed constant $C_\kappa > 0$. Here F and F' are the harmonic extensions of the boundary data f and f', respectively.

At this stage, we make use of Proposition 4.8 (or, more precisely, its equivariant version) to rephrase (4-20) as a fixed-point problem in $e^{\delta s}\mathscr{C}^{2,\alpha}(\mathbb{R} \times S^1)$. The estimates we have just derived are precisely enough to solve this nonlinear problem using a fixed-point argument for contraction mappings in the ball of radius $C_\kappa \tau^2$ in $e^{\delta s}\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)$, where C_κ is a constant which is fixed large enough. Therefore, for all $\tau > 0$ small enough, we find a solution w of (4-20) satisfying

$$\|e^{-\delta s} w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le C_\kappa \tau^2.$$

In addition, it follows from the above estimates that

$$\|e^{-\delta s}(w' - w)\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le C_\kappa \tau^{1/2} \|f' - f\|_{\mathscr{C}^{2,\alpha}(S^1)},$$

where w (respectively w') is the solution associated to f (respectively f').

To complete the result, it is enough to change coordinates $r = \phi(s)$ in the range where $\frac{1}{2}\tau^{3/4} \le r \le 2\tau^{3/4}$ and $|s - \bar{s}| \le 1$. There is no real difficulty in deriving the estimates (4-17) and (4-18), which follow from Proposition 4.7 and the estimate for w. There is a subtlety here: if we change variables $r = \phi(s)$ for $\frac{1}{2}\tau^{3/4} \le r \le 2\tau^{3/4}$ and $|s - \bar{s}| \le 1$, then $F(s)$ is not equal to $W_f^{\text{ins}}(\phi(s))$, because s does not correspond to the cylindrical coordinate $r = e^t$ in $\mathbb{R}^2 - \{0\}$! In fact, we have

$$F(s) = W_f^{\text{ins}}(e^{\bar{s}-s}),$$

and $\tau^{-3/4} r = \phi(s)/\phi(\bar{s})$ and is not equal to $e^{\bar{s}-s}$. Nevertheless, using the expansion of ϕ we have derived, we easily check that

$$\|F(\bar{s} - \log\phi(s) + \log\phi(\bar{s})) - F(s)\|_{\mathscr{C}^{2,\alpha}([\bar{s},\bar{s}+2]\times S^1)} \le c\tau^{1/2}\|f\|_{\mathscr{C}^{2,\alpha}(S^1)}$$

for some constant $c > 0$ independent of τ. □

5. The catenoid

Parametrization and notations. We recall some well-known facts about catenoids in Euclidean space. The normalized catenoid $\mathfrak{C}$ is the minimal surface of revolution, parametrized by

$$Y_0(s,\theta) := (\cosh s \cos\theta, \cosh s \sin\theta, s),$$

where $(s,\theta) \in \mathbb{R}\times S^1$. The induced metric on $\mathfrak{C}$ is given by

$$g_0 := (\cosh s)^2 (ds^2 + d\theta^2),$$

and it is easy to check that the second fundamental form is given by

$$h_0 := ds^2 - d\theta^2$$

when the unit normal vector field is chosen to be

$$N_0 := \frac{1}{\cosh s}(\cos\theta, \sin\theta, -\sinh s).$$

In particular, the formulae for the induced metric and the second fundamental form imply that the mean curvature of the surface $\mathfrak{C}$ is constant and equal to 0.

In the above defined coordinates, the Jacobi operator about the catenoid is given by

$$J_0 := \frac{1}{2\cosh^2 s}\left(\partial_s^2 + \partial_\theta^2 + \frac{2}{\cosh^2 s}\right).$$

Refined asymptotics. We are interested in the parametrization of the catenoid as a (multivalued) vertical graph over the horizontal plane. We consider, for example, the lower part of the catenoid as the graph over the complement of the unit disc in the horizontal plane for the function u_0. Namely, u_0 is the negative function defined by

$$u_0(\cosh s) = s$$

for all $s \le 0$. It is easy to check that

$$s = -\log(2r) + \mathbb{O}_{\mathring{\mathscr{C}}^\infty}(r^{-2}),$$

and hence the lower end of the catenoid can also be parametrized as a vertical graph over $\mathbb{R}^2 - D(0, 1)$ by

$$(r, \theta) \mapsto (r\cos\theta, r\sin\theta, u_0(r)).$$

With little work, one proves the following.

Lemma 5.1. *The expansion*

$$u_0(r) = -\log(2r) + \mathbb{O}_{\mathring{\mathscr{C}}^\infty}(r^{-2})$$

holds in $\mathbb{R}^2 - D(0, 2)$.

Mapping properties of the Jacobi operator about the catenoid. The functional analysis of the Jacobi operator about the catenoid is well understood, and some results can be found, for example, in [Mazzeo et al. 2001]. Again the indicial roots of J_0 characterize the asymptotic behavior of the solutions of the homogeneous problem $J_{0,j} w = 0$, where

$$J_{0,j} := \frac{1}{2\cosh^2 s}\left(\partial_s^2 - j^2 + \frac{2}{\cosh^2 s}\right).$$

It is easy to see that the indicial roots of $J_{0,j}$ are equal to $\pm j$, and hence the indicial roots of J_0 are equal to $\pm j$ for $j \in \mathbb{N}$.

For all $\delta \in \mathbb{R}$, we define the operator

$$\begin{aligned} \mathscr{L}_\delta : (\cosh s)^\delta \mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1) &\to (\cosh s)^\delta \mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1),\\ w &\mapsto (\cosh s)^2 J_0 w, \end{aligned}$$

where, as usual, the norms in the function spaces $\mathscr{C}^{k,\alpha}(\mathbb{R} \times S^1)$ are computed with respect to the cylindrical metric g_{cyl}.

Paralleling what we have proven in Section 4, we have the following.

Proposition 5.2. *Assume that $\delta \in (-2, 2)$. Then there exists $C > 0$, only depending on δ, such that, for all $\bar{w} \in (\cosh s)^\delta \mathscr{C}^{2,\alpha}(\mathbb{R} \times S^1)$, we have*

$$\|(\cosh s)^{-\delta} \bar{w}\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \leq C \|(\cosh s)^{-\delta} \mathscr{L}_\delta \bar{w}\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)},$$

provided

$$\int_{S^1} \bar{w}(s,\theta)\, d\theta = \int_{S^1} \bar{w}(s,\theta) e^{\pm i\theta}\, d\theta = 0 \tag{5-1}$$

for all $s \in \mathbb{R}$.

Proof. The proof is parallel to that of Proposition 4.7 and is left to the reader. □

The following result follows from the general theory developed in [Pacard 2008] (see Theorem 10.2.1 on page 61 and Proposition 12.2.1 on page 81) or [Lockhart and McOwen 1985; Melrose 1993; Mazzeo 1991]. For the sake of completeness we provide a self-contained proof.

Theorem 5.3. *Assume $\delta \in (1, 2)$. Then $\mathscr{L}_\delta$ is surjective and has a 6-dimensional kernel.*

Proof. The proof is similar to that of Proposition 4.8. Recall that the action of rigid motions and dilations provides many Jacobi fields. For example,

$$J_0(\tanh s) = 0 \quad \text{and} \quad J_0(1 - s \tanh s) = 0, \tag{5-2}$$

which either follow from direct computation or from the fact that these are the Jacobi fields associated to the group of vertical translations and the group of dilations centered at the origin.

Similarly

$$J_0\left(\frac{1}{\cosh s} e^{\pm i\theta}\right) = 0 \quad \text{and} \quad J_0\left(\left(\sinh s + \frac{1}{\cosh s}\right) e^{\pm i\theta}\right) = 0, \tag{5-3}$$

which again either follows from direct computation or from the fact that these are the Jacobi fields associated to the group of horizontal translations and the group of rotations about the vertical axis, centered at the origin.

This already shows that when $\delta \in (1, 2)$ the kernel of $\mathscr{L}_\delta$ is at least 6-dimensional.

We first assume that $f \in \mathscr{C}^{0,\alpha}(\mathbb{R} \times S^1)$ has compact support and we decompose it as $f(s,\theta) = f_0(s) + f_{\pm 1}(s) e^{\pm i\theta} + \bar{f}(s,\theta)$, where, by definition,

$$\bar{f} := \sum_{j \neq 0, \pm 1} f_j(s) e^{ij\theta}.$$

If we restrict our attention to functions $\bar{w}$ whose Fourier decomposition in the θ variable is of the form

$$\bar{w}(s,\theta) = \sum_{j \neq 0, \pm 1} w_j(s) e^{ij\theta},$$

we have

$$\int_{\mathbb{R}\times S^1} \left(|\partial_s \bar{w}|^2 + |\partial_\theta \bar{w}|^2 - \frac{2}{\cosh^2 s} \bar{w}^2 \right) ds\, d\theta \geq 2 \int_{\mathbb{R}\times S^1} \bar{w}^2 \, ds\, d\theta. \tag{5-4}$$

Therefore, we can solve

$$(\cosh s)^2 J_0 \bar{w} = \bar{f}$$

in $H^1(\mathbb{R}\times S^1)$. Elliptic estimates then imply that $\bar{w} \in \mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)$.

Obviously $\bar{w} \in (\cosh s)^\delta \mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)$, since $\delta > 0$. Proposition 5.2 applies, and we get

$$\|(\cosh s)^{-\delta} \bar{w}\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \leq C \|(\cosh s)^{-\delta} f\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)}$$

for any function f having compact support. The general result, when f does not necessarily have compact support, follows from a standard exhaustion argument.

Finally, the solvability of

$$(\cosh s)^2 J_0(w_j e^{ij\theta}) = f_j e^{ij\theta}$$

for $j = 0, \pm 1$, follows easily from integration of the associated second order ordinary differential equation starting from 0 (with initial data and initial velocity equal to 0). We have, explicitly,

$$w_j = A_j^+ \int_0^s A_j^-(t) f_j(t)\, dt - A_j^- \int_0^s A_j^+(t) f_j(t)\, dt,$$

where $A_j^\pm$ are the two independent solutions of

$$\left(\partial_s^2 - j^2 + \frac{2}{\cosh^2 s} \right) A_j^\pm = 0,$$

which are given in (5-2) and (5-3) and are normalized so that their Wronskian is equal to 1. Direct estimates imply that

$$\|(\cosh s)^{-\delta} w_j\|_{\mathscr{C}^{2,\alpha}(\mathbb{R})} \leq C \|(\cosh s)^{-\delta} f\|_{\mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1)},$$

provided $\delta > 1$ (more precisely, $\delta > 0$ is needed to derive the estimate for w_0 and $\delta > 1$ is needed to derive the estimate for $w_{\pm 1}$). We set $w = w_0 + w_{\pm 1} e^{\pm i\theta} + \bar{w}$. This completes the proof of the fact that the operator $\mathscr{L}_\delta$ is surjective when $\delta \in (1,2)$. The fact that this operator, restricted to the space of functions satisfying the orthogonality conditions (5-1), is injective follows from Proposition 5.2. Hence the kernel of $\mathscr{L}_\delta$ is 6-dimensional. $\square$

The mean curvature of normal graphs over the catenoid. We consider in this section the mean curvature of a surface which is a normal graph over $\mathfrak{C}$. Hence, for some smooth (small enough) function w defined on $\mathfrak{C}$, we consider the surface parametrized by

$$Y(s,\theta) = Y_0(s,\theta) + w(s,\theta) N_0(s,\theta).$$

We have the following technical result.

Lemma 5.4. *The mean curvature of the surface parametrized by Y is given by*

$$H(w) = J_0 w + \frac{1}{\cosh s} Q_0\left(\frac{w}{\cosh s}\right),$$

where the nonlinear second order differential operator Q_0 satisfies

$$\|Q_0(v_2) - Q_0(v_1)\|_{\mathscr{C}^{0,\alpha}([s,s+1]\times S^1)} \\ \leq c(\|v_1\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)} + \|v_2\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)}) \|v_2 - v_1\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)}$$

for some constant $c > 0$ independent of s and $\tau \in (0, 1)$ and for all functions v_1, v_2 satisfying $\|v_i\|_{\mathscr{C}^{2,\alpha}([s,s+1]\times S^1)} \leq 1$.

Proof. This result is already proven in [Mazzeo and Pacard 2001]. In any case, a simple proof follows easily from Proposition 2.1 together with the fact that

$$g_w = \cosh^2 s\left(\left(1 - \frac{w}{\cosh^2 s}\right)^2 ds^2 + \left(1 + \frac{w}{\cosh^2 s}\right)^2 d\theta^2\right).$$

We leave the details to the reader. □

A second fixed-point argument. Assume that $\tau, \tilde{\tau} > 0$ are chosen small enough and satisfy

$$\left|\tilde{\tau} - \frac{\tau}{m+1}\right| \leq \kappa \tau^{3/2}, \tag{5-5}$$

where the constant $\kappa > 0$ is large enough; its value will be fixed in Section 7. The rationale for this estimate is also explained in Section 7. We define $\tilde{s} > 0$ by

$$\tilde{\tau} \cosh \tilde{s} = \tau^{3/4}.$$

Observe that $\tilde{s}$ depends on both τ and $\tilde{\tau}$ even though we have chosen not to make this apparent in the notation. It is easy to check that $\tilde{s} = -(1/4)\log \tau + \mathbb{O}(1)$. We define the truncated catenoid $\mathfrak{C}_{\tilde{\tau}}$ to be the image of $[-\tilde{s}, \tilde{s}] \times S^1$ by $\tilde{\tau} Y_0$ (to simplify the notations, we do not write the dependence of this surface on the parameter τ).

Building on the previous analysis, we prove the existence of *constant mean curvature surfaces* which are close to the truncated catenoid $\mathfrak{C}_{\tilde{\tau}}$ and have two boundaries which can be described by some function $f : S^1 \to \mathbb{R}$. We also require that the surfaces are invariant under the action of the symmetry with respect to the horizontal plane. More precisely, we have the following.

Proposition 5.5. *Assume we are given $\kappa > 0$ large enough (to be fixed later). For all τ, $\tilde{\tau} > 0$ small enough satisfying (5-5) and for all functions f invariant under the action of the $\mathrm{Dih}^{(2)}_{m+1}$ satisfying (H1) (notice that (H2) is automatically satisfied) and*

$$\|f\|_{\mathscr{C}^{2,\alpha}(S^1)} \leq \kappa \tau^{3/2}, \tag{5-6}$$

there exists a constant mean curvature 1 surface $\mathfrak{C}_{\tilde{\tau},f}$ which is close to $\mathfrak{C}_{\tilde{\tau}}$ and has two boundaries. The surface $\mathfrak{C}_{\tilde{\tau},f}$ is invariant under the action of $\mathscr{S}_3$, the symmetry with respect to the horizontal plane $x_3 = 0$, $\mathscr{S}_2$, the symmetry with respect to the plane $x_2 = 0$. And, close to its lower boundary, the surface $\mathfrak{C}_{\tilde{\tau},f}$ is a vertical graph *over the annulus*

$$\{x \in \mathbb{R}^2 : \tfrac{1}{2}\tau^{3/4} \leq |x| \leq \tau^{3/4}\},$$

for some function $x \mapsto U^{\downharpoonleft}_{\tilde{\tau},f}(\tau^{-3/4}x)$ which can be expanded as

$$U^{\downharpoonleft}_{\tilde{\tau},f}(x) = -\tilde{\tau}\log\frac{2\tau^{3/4}}{\tilde{\tau}} - \tilde{\tau}\log|x| + W^{\mathrm{ins}}_f(x) + \overline{U}^{\downharpoonleft}_{\tilde{\tau},f}(x), \tag{5-7}$$

where we recall that W^{ins}_f denotes the bounded harmonic extension of f in the punctured unit disc and where

$$\|\overline{U}^{\downharpoonleft}_{\tilde{\tau},0}\|_{\mathscr{C}^{2,\alpha}\left(\bar{D}(0,1)-D\left(0,\frac{1}{2}\right)\right)} \leq C\tau^{3/2}. \tag{5-8}$$

Moreover, the nonlinear mapping

$$\mathscr{C}^{2,\alpha}(S^1) \ni f \mapsto \overline{U}^{\downharpoonleft}_{\tilde{\tau},f} \in \mathscr{C}^{2,\alpha}\left(\bar{D}(0,1) - D\left(0,\tfrac{1}{2}\right)\right)$$

is Lipschitz, and, given $\delta \in (1,2)$, we have

$$\|\overline{U}^{\downharpoonleft}_{\tilde{\tau},f'} - \overline{U}^{\downharpoonleft}_{\tilde{\tau},f}\|_{\mathscr{C}^{2,\alpha}\left(\bar{D}(0,1)-D\left(0,\frac{1}{2}\right)\right)} \leq C\tau^{(2-\delta)/4}\|f'-f\|_{\mathscr{C}^{2,\alpha}(S^1)} \tag{5-9}$$

for some constant $C > 0$ independent of κ, τ, $\tilde{\tau}$ and f, f'. The function $\overline{U}^{\downharpoonleft}_{\tilde{\tau},f}$ depends continuously on $\tilde{\tau}$.

Proof. The proof of this result is very similar to the proof of Proposition 4.12, so we cover only the main differences.

Again, in the annular region which is the image of $(-\tilde{s}-2, -\tilde{s}+2) \times S^1$ by Y_0, we modify the unit vector field N_0 into $\bar{N}_0$ in such a way that $\bar{N}_0$ is equal to e_3 on the image of $(-\tilde{s}-1, \tilde{s}+1) \times S^1$ by $\tilde{\tau}Y_0$. We perform a similar modification on the upper half of the catenoid, on the image of $(\tilde{s}-2, \tilde{s}+2) \times S^1$ by Y_0, so that our construction remains invariant under the action of the symmetry with respect to the horizontal plane. In this case, using Proposition 2.2, one can check that the expression of the mean curvature given in Lemma 4.11 has to be altered to

$$\overline{H}(w) = J_0 w + \frac{1}{\cosh^2 s} l_0 w + \frac{1}{\cosh s}\,\overline{Q}_0\Big(\frac{w}{\cosh s}\Big),$$

where $\overline{Q}_0$ enjoys properties which are similar to the properties enjoyed by Q_0 and where l_0 is a linear second order partial differential operator in ∂_s and ∂_θ whose coefficients are smooth, supported in $(-\tilde{s}-2, -\tilde{s}+2)\times S^1$ and in $(\tilde{s}-2, \tilde{s}+2)\times S^1$, and are bounded (in the smooth topology) by a constant, independent of τ, times $\tau^{1/2}$.

We assume that f is chosen to satisfy (H1), (H2), and

$$\|f\|_{\mathscr{C}^{2,\alpha}(S^1)} \leq \kappa\tau^{1/2}.$$

Observe that the norm of the boundary data f is bounded by a constant times $\tau^{1/2}$ and not $\tau^{3/2}$, the reason being that we are going to perturb the image of $[-\tilde{s}, \tilde{s}]\times S^1$ by Y_0 and then scale the surface we obtain by a factor $\tilde{\tau}$ instead of perturbing $\mathfrak{C}_{\tilde{\tau}}$, which is the image of $[-\tilde{s}, \tilde{s}]\times S^1$ by $\tilde{Y}_0$. This is also the reason the equation we solve is $\overline{H}(w) = \tilde{\tau}$ and not $\overline{H}(w) = 1$.

We denote by F the harmonic extension of f in $(-\infty, \tilde{s})\times S^1$ and we set

$$\tilde{F}(s,\theta) := F(s,\theta) + F(-s,\theta),$$

which is well defined in $[-\tilde{s}, \tilde{s}]\times S^1$. One should be aware that the boundary data of F is not exactly equal to f, but the difference between F and f on the boundary tends to 0 as τ tends to 0. More precisely, we have

$$\|F-\tilde{F}\|_{\mathscr{C}^{2,\alpha}([\tilde{s}-1,\tilde{s}]\times S^1)} \leq C\tau\|f\|_{\mathscr{C}^{2,\alpha}(S^1)}.$$

We would like to solve the equation

$$(\cosh s)^2 J_0(\tilde{F}+w) + l_0(\tilde{F}+w) + \cosh s\, Q_0\Big(\frac{\tilde{F}+w}{\cosh s}\Big) = (\cosh s)^2\tilde{\tau} \tag{5-10}$$

in $(-\tilde{s}, \tilde{s})\times S^1$. This will provide constant mean curvature surfaces with mean curvature equal to $\tilde{\tau}$ which are close to the truncated catenoid. Again, the solvability of this nonlinear problem follows from a fixed-point theorem for a contraction mapping.

We choose

$$\bar{\mathscr{E}}_\tau : \mathscr{C}^{0,\alpha}([-\tilde{s}, \tilde{s}]\times S^1) \to \mathscr{C}^{0,\alpha}(\mathbb{R}\times S^1),$$

an extension operator, such that

$$\begin{cases} \bar{\mathscr{E}}_\tau(\psi) = \psi & \text{in } [-\tilde{s}, \tilde{s}]\times S^1, \\ \bar{\mathscr{E}}_\tau(\psi) = 0 & \text{in } ((-\infty, -\tilde{s}-1]\cup[\tilde{s}+1, \infty))\times S^1, \end{cases}$$

and

$$\|\bar{\mathscr{E}}_\tau(\psi)\|_{\mathscr{C}^{0,\alpha}([-\tilde{s}-1,-\tilde{s}+1]\times S^1)} \le C\|\psi\|_{\mathscr{C}^{0,\alpha}([-\tilde{s},-\tilde{s}+1]\times S^1)},$$
$$\|\bar{\mathscr{E}}_\tau(\psi)\|_{\mathscr{C}^{0,\alpha}([\tilde{s}-1,\tilde{s}+1]\times S^1)} \le C\|\psi\|_{\mathscr{C}^{0,\alpha}([\tilde{s}-1,\tilde{s}']\times S^1)}.$$

We rewrite (5-10) as

$$(5\text{-}11)\quad (\cosh s)^2 J_0 w = \bar{\mathscr{E}}_\tau\Big((\cosh s)^2(\tilde{\tau}-J_0\tilde{F}) - l_0(\tilde{F}+w) - \cosh s\, Q_0\Big(\frac{\tilde{F}+w}{\cosh s}\Big)\Big).$$

Again, on the right side it is understood that we consider the image by $\bar{\mathscr{E}}_\tau$ of the restriction of the functions to $[-\tilde{s},\tilde{s}]\times S^1$.

We assume that $\delta \in (1,2)$ is fixed. It is easy to check that there exists a constant $c>0$ (independent of κ) and a constant $c_\kappa>0$ (depending on κ) such that

$$\|(\cosh s)^{-\delta}\bar{\mathscr{E}}_\tau(\cosh^2 s\,\tilde{\tau})\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le c\tau^{(2+\delta)/4},$$
$$\Big\|(\cosh s)^{-\delta}\bar{\mathscr{E}}_\tau\Big(\frac{\tilde{F}}{\cosh^2 s}\Big)\Big\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le c\tau^{1/2}\|f\|_{\mathscr{C}^{2,\alpha}(S^1)},$$
$$\|(\cosh s)^{-\delta}\bar{\mathscr{E}}_\tau(l_0(\tilde{F}+w))\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)}$$
$$\le c\tau^{1/2}(\|f\|_{\mathscr{C}^{2,\alpha}(S^1)} + \tau^{\delta/4}\|(\cosh s)^{-\delta}w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)},),$$

and

$$\Big\|(\cosh s)^{-\delta}\bar{\mathscr{E}}_\tau\Big(\cosh s\, Q\Big(\frac{w'+\tilde{F}'}{\cosh s}\Big) - \cosh s\, Q\Big(\frac{w+\tilde{F}}{\cosh s}\Big)\Big)\Big\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)}$$
$$\le c_\kappa(\tau^{3/4}\|(\cosh s)^{-\delta}(w'-w)\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} + \tau^{(1+\delta)/4}\|f'-f\|_{\mathscr{C}^{2,\alpha}(S^1)}),$$

provided w and w' satisfy

$$\|(\cosh s)^{-\delta}w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} + \|(\cosh s)^{-\delta}w'\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le C\tau^{(2+\delta)/4}$$

for some fixed constant $C>0$ (independent of κ, τ, and f). Here $\tilde{F}$ and $\tilde{F}'$ are the harmonic extensions of the boundary data f and f', respectively.

Now we make use of Theorem 5.3 to rephrase the problem as a fixed-point problem, and the previous estimates are precisely enough to solve this nonlinear problem using a fixed-point argument for contraction mappings in the ball of radius $C\tau^{(2+\delta)/4}$ in $(\cosh s)^\delta\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)$, where $C>0$ is fixed large enough independent of κ, provided τ is small enough. Then, for all $\tau>0$ small enough, we find that there exists constants $C>0$ (independent of κ) and $C_\kappa>0$ (depending on κ) such that, for all functions f satisfying (H1), (H2), and (4-15), there exists a unique solution w of (4-20) satisfying

$$\|(\cosh s)^\delta w\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le C\tau^{(2+\delta)/4}.$$

In addition, we have the estimate

$$\|e^{-\delta s}(w' - w)\|_{\mathscr{C}^{2,\alpha}(\mathbb{R}\times S^1)} \le C_\kappa \tau^{1/2} \|f' - f\|_{\mathscr{C}^{2,\alpha}(S^1)},$$

where w (respectively w') is the solution associated to f (respectively f').

To complete the result, we simply shrink the surface we have obtained by a factor $\tilde{\tau}$ to get a surface whose mean curvature is constant and equal to 1. The description of this surface close to its boundary follows from the arguments already used in the proof of Proposition 4.12. Observe that the solution of (4-20) is obtained through a fixed-point theorem for contraction mappings, and it is classical to check that the solution we obtain depends continuously on the parameters of the construction. In particular, the constant mean curvature surface we obtain depends continuously on $\tilde{\tau}$ (in fact one can also prove that the surface depends smoothly on $\tilde{\tau}$, but we shall not use this property).

To prove that, near its lower boundary, the surface we have obtained is a vertical graph for some function which enjoys the decomposition (5-7), we make use of the expansion in Lemma 5.1, and we follow the steps of the construction. Notice that $\overline{U}^{\downharpoonleft}_{\tilde{\tau},f}$ collects many remainders: the one coming from the expansion in Lemma 5.1, the difference between F and $\tilde{F}$, the solution w of the fixed-point problem. Also, the change of coordinates which takes into account that the variable s does not correspond to the cylindrical coordinates in $\mathbb{R}^2 - \{0\}$. □

6. The unit sphere

Notations and definitions. We denote by x_1, x_1, x_3 the coordinates in $\mathbb{R}^3$, and by $\mathscr{S}_j$ the symmetry with respect to the plane $x_j = 0$. For all $m \in \mathbb{N}$, we write $\mathscr{R}_{m+1}$ for a rotation of angle $2\pi/(m+1)$ about the x_3-axis. With slight abuse of notation, we will keep the same notation to denote the restriction of these isometries to the horizontal plane.

We define $z_0, \dots, z_m \in S^1$ to be vertices of a regular polygon with $m+1$ edges in the plane. Without loss of generality, we can choose

$$z_0 := (1, 0) = e_1 \in \mathbb{R}^2,$$

and, for $j = 1, \dots, m-1$, $z_{j+1} \in \mathbb{R}^2$ is the image of z_j by $\mathscr{R}_{m+1}$. Thus, if we identify the horizontal plane with $\mathbb{C}$, the vertices of the polygon are exactly the $(m+1)$-th roots of unity. Recall that the dihedral group of symmetries of $\mathbb{R}^2$ leaving this polygon fixed has been denoted by $\mathrm{Dih}^{(2)}_{m+1}$. It is generated by $\mathscr{R}_{m+1}$ and $\mathscr{S}_2$.

Let S^2 be the unit sphere in $\mathbb{R}^3$. The upper half hemisphere of S^2 can be parametrized by

$$X^{\upharpoonleft}(x) := (x, \sqrt{1 - |x|^2}),$$

while the lower hemisphere is parametrized by

$$X^{\downharpoonleft}(x) := (x, -\sqrt{1-|x|^2}),$$

where, in both cases, $x \in D(0, 1)$.

For all $\tau > 0$ small enough, we set

$$B^{\upharpoonleft} := X^{\upharpoonleft}(D(0, \tau^{3/4})),$$

and, for all $\rho > 0$ satisfying

$$\frac{1}{C}\tau \le \rho^2 \le C\tau, \tag{6-1}$$

for some fixed constant $C > 1$, we define

$$B_j^{\downharpoonleft} := X^{\downharpoonleft}(D(\rho z_j, \tau^{3/4}))$$

for $j = 0, \dots, m$. We also define

$$p^{\upharpoonleft} := X^{\upharpoonleft}(0)$$

to be the north pole of S^2, and, for $j = 0, \dots, m$, we define the points

$$p_j^{\downharpoonleft} := X^{\downharpoonleft}(\rho z_j),$$

which are $m+1$ points arranged symmetrically near the south pole of S^2. By construction, $p_{j+1}^{\downharpoonleft}$ is the image of $p_j^{\downharpoonleft}$ by $\mathcal{R}_{m+1}$.

Definition 6.1. We define $\mathfrak{S}_{\tau,\rho}$ to be the surface obtained by excising from S^2 the sets $B^{\upharpoonleft}$ and $B_j^{\downharpoonleft}$ for $j = 0, \dots, m$.

Observe that, provided τ is chosen small enough, the surface $\mathfrak{S}_{\tau,\rho}$ has $m+2$ boundaries. Moreover, this surface has been constructed in such a way that it is invariant under the action of the dihedral group $\mathrm{Dih}_{m+1}^{(2)}$.

The mean curvature of vertical graphs. We recall some well-known facts about the mean curvature of vertical graphs in $\mathbb{R}^3$. The mean curvature of the graph of the function $u \in \mathscr{C}^2_{\mathrm{loc}}(\mathbb{R}^2)$, namely, the surface parametrized by

$$X(x) := (x, u(x)) \in \mathbb{R}^3,$$

where x belongs to some open domain in $\mathbb{R}^2$, is given by

$$M(u) := \tfrac{1}{2}\,\mathrm{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right).$$

Recall that the mean curvature is defined to be the average of the principal curvatures; this explains the factor $\frac{1}{2}$.

It follows from this formula that the linearized mean curvature operator about the graph of u is given by

$$DM(u)(v) = \frac{\Delta v}{2W} + \frac{3}{2W^5}(\nabla u \cdot \nabla v) D^2 u[\nabla u, \nabla u] \\ - \frac{1}{2W^3}\big((\nabla u \cdot \nabla v)\Delta u + D^2 v[\nabla u, \nabla u] + 2D^2 u[\nabla u, \nabla v]\big),$$

where

$$W := \sqrt{1+|\nabla u|^2},$$

and where $D^2 f[\,\cdot\,,\cdot\,]$ is the second order differential of the function f. One should be aware that $DM(u)$ is not the Jacobi operator J_u about the graph of the function u, since nearby surfaces are not parametrized as normal graphs, but as vertical graphs over the horizontal plane. As explained in Section 2, this operator and the Jacobi operator are conjugate, and, in fact, assuming the vertical graph is oriented so that the unit normal vector field points upward, we have the relation

$$X^*(J_u w) = DM(u)(W X^* w)$$

for any function defined on the graph of u.

Of interest is the case where, for example,

$$u(x) = \pm\sqrt{1-|x|^2},$$

where $x = (x_1, x_2) \in \mathbb{R}^2$. According to the sign chosen, the graph of u is the lower or the upper hemisphere of the sphere of radius 1 centered at the origin. In this case, we have

$$\nabla u(x) = \mp \frac{x}{\sqrt{1-|x|^2}}, \qquad \nabla^2 u(x) = \mp \frac{(1-|x|^2)\,\mathrm{Id} + x \otimes x}{(1-|x|^2)^{3/2}}$$

and

$$\Delta u(x) = \mp \frac{2-|x|^2}{(1-|x|^2)^{3/2}}.$$

Using these, we find that the explicit expression of $DH(u)$ in $D(0,1)$ is

$$DM(u)w = \tfrac{1}{2}(1-|x|^2)^{1/2}(\Delta w - \nabla^2 w(x,x) - 4(x \cdot \nabla w)). \tag{6-2}$$

Green's function. Let N_0 denote the inward pointing unit normal vector field on S^2. We consider an inward pointing vector field $N_0^\flat$, which is equal to N_0 close to the (horizontal) equator of S^2 and which is equal to a vertical unit vector field close to the north and south pole of S^2 (still pointing inward).

We define $\mathbb{L}$ to be the linearized mean curvature operator using the vector field $N_0^\flat$. According to the analysis in Section 2, we can write

$$\mathbb{L}w := \tfrac{1}{2}(\Delta_{S^2}+2)(N_0 \cdot N_0^\flat w). \tag{6-3}$$

We let Γ_ρ be the unique solution of

$$\mathbb{L}\Gamma_\rho = -\pi\delta_{p^\upharpoonleft} - \frac{\pi}{\sqrt{1-\rho^2}}\frac{1}{m+1}(\delta_{p_0^\downharpoonleft}+\cdots+\delta_{p_m^\downharpoonleft}), \tag{6-4}$$

which satisfies the orthogonality conditions

$$\int_{S^2} x_i \Gamma_\rho \, d\mathrm{vol}_{S^2} = 0$$

for $i = 1, 2$ and 3. Here δ_q is the Dirac mass at the point q. The existence of Γ_ρ is guaranteed by the fact that the distribution on the right side of (6-4) is orthogonal to the cokernel of $\mathbb{L}$. Indeed, the Jacobi operator is selfadjoint and its kernel and cokernel are equal and spanned by the restriction of the coordinate functions to the unit sphere. Thanks to (6-3), we conclude that the cokernel of $\mathbb{L}$ is also spanned by the restriction of the coordinate functions to the unit sphere, multiplied by the factor $N_0 \cdot N_0^\flat$. For the sake of simplicity, we define the functions

$$\tilde{x}_i := x_i N_0 \cdot N_0^\flat$$

for $i = 1, 2, 3$, which are obtained by multiplication of the coordinate functions by $N_0 \cdot N_0^\flat$.

Now

$$\langle x_1, \delta_{p^\upharpoonleft}\rangle_{\mathscr{D},\mathscr{D}'} = \langle x_2, \delta_{p^\upharpoonleft}\rangle_{\mathscr{D},\mathscr{D}'} = 0,$$

since both x_1 and x_2 vanish at the north pole of S^2 and

$$\langle x_1, \delta_{p_j^\downharpoonleft}\rangle_{\mathscr{D},\mathscr{D}'} = \rho\cos\Big(\frac{2\pi}{m+1}j\Big) \quad \text{and} \quad \langle x_2, \delta_{p_j^\downharpoonleft}\rangle_{\mathscr{D},\mathscr{D}'} = \rho\sin\Big(\frac{2\pi}{m+1}j\Big).$$

Since

$$\sum_{j=0}^{m}\cos\Big(\frac{2\pi}{m+1}j\Big) = \sum_{j=0}^{m}\sin\Big(\frac{2\pi}{m+1}j\Big) = 0,$$

we conclude that the distribution on the right side of (6-4) is orthogonal to the coordinate functions x_1 and x_2. Geometrically, this just follows from the fact that the points $p_j^\downharpoonleft$ are symmetrically arranged around the x_3-axis. Finally, we have

$$\langle x_3, \delta_{p^\upharpoonleft}\rangle_{\mathscr{D},\mathscr{D}'} = 1 \quad \text{and} \quad \langle x_3, \delta_{p_j^\downharpoonleft}\rangle_{\mathscr{D},\mathscr{D}'} = -\sqrt{1-\rho^2},$$

and, again, we conclude that the distribution on the right side of (6-4) is orthogonal to the coordinate function x_3 thanks to the choice of the constant in front of the Dirac masses at the points $p_j^\natural$. Geometrically, this has some interesting consequences and can be interpreted as a conservation of the vertical flux of the surfaces we try to construct. We shall return to this point later on. Finally, observe that Γ_ρ is invariant under the action of the elements of $\mathrm{Dih}^{(2)}_{m+1}$.

We now turn to the expansion of the function Γ_ρ near the north pole of S^2.

Lemma 6.2. *The expansion*

$$X^{\natural *}\Gamma_\rho(x) = -\log|x| + a^\natural + \mathcal{O}_{\mathcal{C}^\infty}(|x|^2)$$

holds in a fixed neighborhood of 0, *where the constant $a^\natural \in \mathbb{R}$ depends smoothly on ρ and is bounded as ρ tends to* 0. *Moreover*, *the estimate on $\mathcal{O}_{\mathcal{C}^\infty}(|x|^2)$ is uniform as ρ tends to* 0.

Proof. We define the function Γ_0 on the upper hemisphere by

$$X^{\natural *}\Gamma_0(x) = -\log|x|,$$

and, using (6-2), we compute

$$X^{\natural *}(\mathbb{L}\Gamma_0 + \pi\delta_{p^\natural}) = \tfrac{3}{2}\sqrt{1-|x|^2}.$$

This immediately implies that, close to $p^\natural$, the function $\Gamma_\rho - \Gamma_0$ is smooth. In particular, this function can be expanded as

$$X^{\natural *}(\Gamma_\rho - \Gamma_0)(x) = a^\natural + b^\natural \cdot x + \mathcal{O}_{\mathcal{C}^\infty}(|x|^2),$$

where $a^\natural \in \mathbb{R}$ and $b^\natural \in \mathbb{R}^2$ depend smoothly on ρ and remain bounded as ρ tends to 0. Since the function Γ_ρ is also invariant under the action of the elements of $\mathrm{Dih}^{(2)}_{m+1}$, we conclude that necessarily $b^\natural = 0$. This completes the proof. □

Near the other poles, the function Γ_ρ also has an expansion which we now describe. As can be suspected, this description relies on the expansion of the function

$$G(x) := -\sum_{j=0}^{m} \log|x - \rho z_j|$$

at any of its singularities. Since this is a key point in our construction, we spend some time deriving this expansion carefully. By symmetry, it is enough to expand this function at ρz_0. We change variables and write $x = \rho z_0 + y$. We then expand

$$\log|y - \rho(z_j - z_0)| = \log\rho + \log|z_j - z_0| + \frac{1}{\rho}\,\frac{z_0 - z_j}{|z_0 - z_j|^2}\cdot y + \mathcal{O}\left(\frac{|y|^2}{\rho^2}\right).$$

Hence we find

$$G(\rho z_0 + y) = -\log|y| - m\log\rho - \sum_{j=1}^{m}\log|z_j - z_0| - \frac{1}{\rho}\sum_{j=1}^{m}\frac{z_0 - z_j}{|z_0 - z_j|^2}\cdot y + \mathcal{O}\left(\frac{|y|^2}{\rho^2}\right).$$

It is easy to check the identity

$$\sum_{j=1}^{m} \frac{z_0 - z_j}{|z_0 - z_j|^2} = \frac{m}{2} z_0.$$

Setting

$$a_0^{\downharpoonleft} := \sum_{j=1}^{m} \log |z_j - z_0|,$$

we can write

$$G(\rho z_0 + y) = -\log |y| - m \log \rho - a_0^{\downharpoonleft} - \frac{m}{2\rho} z_0 \cdot y + \mathbb{O}\left(\frac{|y|^2}{\rho^2}\right).$$

Similar estimates can be obtained for the partial derivatives of G. Finally, we have

$$\Delta G = -2\pi(\delta_{\rho z_0} + \cdots + \delta_{\rho z_m}).$$

We now prove that, at $p_j^{\downharpoonleft}$, the expansion of the function $X^{\downharpoonleft *}\Gamma_\rho$ is (in some sense to be made precise) close to the expansion of G near ρz_j.

Lemma 6.3. *The expansion*

$$X^{\downharpoonleft *}\Gamma_\rho(\rho z_j + y) = -\frac{1}{m+1}\left(\log |y| + m \log \rho + a_{0,\rho}^{\downharpoonleft} + \frac{m}{2\rho} z_j \cdot y\right) + \mathbb{O}_{\mathscr{C}^\infty}(\tau^{1/2})$$

holds for $|y| \in \left[\frac{1}{2}\tau^{3/4}, 2\tau^{3/4}\right]$. *Here* $a_{0,\rho}^{\downharpoonleft} \in \mathbb{R}$ *smoothly depends on* $\rho > 0$ *and is uniformly bounded as* ρ *tends to* 0.

Proof. Thanks to the invariance with respect to the action of $\mathrm{Dih}_{m+1}^{(2)}$, it is enough to describe this expansion near the point $p_0^{\downharpoonleft}$. As in the proof of Lemma 6.2, we show that, near the south pole of S^2, the function $X^{\downharpoonleft *}\Gamma_\rho$ is not too far from G. To this end, we define $\tilde{\Gamma}_\rho$ on the lower hemisphere of S^2 by

$$X^{\downharpoonleft *}\tilde{\Gamma}_\rho = G,$$

and, thanks to (6-2), we can compute

$$X^{\downharpoonleft *}(\mathbb{L}\tilde{\Gamma}_\rho + \pi\sqrt{1-\rho^2}(\delta_{p_0^{\downharpoonleft}} + \cdots + \delta_{p_m^{\downharpoonleft}}))$$
$$= \tfrac{1}{2}\sqrt{1-|x|^2} \sum_{j=0}^{m} \left(3 - 2\rho \frac{z_j \cdot (x - \rho z_j)}{|x - \rho z_j|^2} + \frac{\rho^2}{|x - \rho z_j|^2}\left(1 - 2\frac{(z_j \cdot (x - \rho z_j))^2}{|x - \rho z_j|^2}\right)\right).$$

Observe that the right side contains three terms which have different regularity properties. The first one is a smooth function which depends smoothly on ρ and which is invariant by rotation. The second function has a singularity of order 1 at each ρz_j and is bounded by a constant times $\rho|x - \rho z_j|^{-1}$. The third function has a singularity of order 2 at each ρz_j and is bounded by a constant times $\rho^2|x - \rho z_j|^{-2}$.

As a consequence, $X^{\downharpoonleft *}(\tilde{\Gamma}_\rho - \Gamma_\rho)$ can be decomposed into the sum of three functions, which can be analyzed independently. The first one, $f_\rho^{(1)}$, is smooth in a fixed neighborhood of 0 and depends smoothly on the parameter ρ. This implies that, near each ρz_j, this function has a Taylor expansion with coefficients smoothly depending on ρ. Hence

$$f_\rho^{(1)}(x) = f_\rho^{(1)}(\rho z_0) + \nabla f_\rho^{(1)}(\rho z_0) \cdot (x - \rho z_0) + \mathbb{O}(|x - \rho z_0|^2).$$

Observe that $\nabla f_\rho^{(1)}(0) = 0$, and hence $|\nabla f_\rho^{(1)}(\rho z_0)| \le C\rho$. We conclude that $f_\rho^{(1)}(x) = f_\rho^{(1)}(\rho z_0) + \mathbb{O}(\tau^{5/4})$ when $|x - \rho z_0| \in \left[\frac{1}{2}\tau^{3/4}, 2\tau^{3/4}\right]$.

Since $\sum_j |z - z_j|^{-1} \in L^p\big(D\big(0, \frac{1}{2}\big)\big)$ for all $p \in (1, 2)$, we find that the second function

$$f_\rho^{(2)} \in W^{2,p}(D(0, 1/3)),$$

and hence it is continuous near ρz_0 and $f_\rho^{(2)}(x) - f_\rho^{(2)}(\rho z_0)$ is bounded by a constant times $\rho \sum_{j=0}^m |x - \rho z_j|^\nu$ for any given $\nu < 1$. In particular, $f_\rho^{(2)}(x) = f_\rho^{(2)}(\rho z_0) + \mathbb{O}(\tau^{(2+3\nu)/4})$ when $|x - \rho z_0| \in \left[\frac{1}{2}\tau^{3/4}, 2\tau^{3/4}\right]$.

Finally, using Proposition 6.6, the third function $f_\rho^{(3)}$ is bounded by a constant times $\rho^2 \sum_{j=0}^m |x - \rho z_j|^\mu$ for any $\mu \in (-1, 0)$.

In particular, when $|x - \rho z_0| \in \left[\frac{1}{2}\tau^{3/4}, 2\tau^{3/4}\right]$, we find that the sum of these functions can be decomposed as the sum of a constant function (smoothly depending on ρ) and a function which is bounded by a constant times $\tau^{1/2}$ (choose $\nu = \frac{1}{2}$ and $\mu = -\frac{1}{2}$). The statement then follows at once. □

It is interesting to observe that Γ_ρ depends on $\rho > 0$ since the points $p_j^{\downharpoonleft}$ do, and, as ρ tends to 0, the sequence Γ_ρ converges on compact sets to the unique solution of

$$\mathbb{L}\Gamma_0 = -\pi(\delta_{p^{\upharpoonleft}} + \delta_{p^{\downharpoonleft}}),$$

which is L^2-orthogonal to the smooth kernel of $\Delta_{S^2} + 2$. Recall that $p^{\upharpoonleft}$ denotes the north pole of S^2, and we now agree that $p^{\downharpoonleft}$ denotes the south pole of S^2.

Remark 6.4. If a solution to $\mathbb{L}w = 0$ is defined in $S^2 - \{p^{\upharpoonleft}, p^{\downharpoonleft}\}$, is invariant under the action of $\mathrm{Dih}_{m+1}^{(2)}$, and is bounded by a constant times $\mathrm{dist}(\,\cdot\,, \{p^{\upharpoonleft}, p^{\downharpoonleft}\})^\nu$ for some $\nu \in (-1, 0)$, then it is a linear combination of $\tilde{x}_3$ and Γ_0. This will be useful later.

We now summarize the above analysis. We set

$$u^{\upharpoonleft}(x) := \sqrt{1 - |x|^2} \quad \text{and} \quad u^{\downharpoonleft}(x) := -\sqrt{1 - |x|^2}.$$

Observe that, thanks to the previous results, near 0, the graph of the function

$$v^{\upharpoonleft} := u^{\upharpoonleft} + \tau X^{\upharpoonleft *}\Gamma_\rho$$

can be expanded as

$$v^{\upharpoonleft}(x) = 1 + \tau(m \log \rho + a^{\upharpoonleft}) + \tau \log |x| + \mathcal{O}_{\mathring{\mathscr{C}}^\infty}(\tau^{3/2})$$

for $|x| \in [(1/2)\tau^{3/4}, 2\tau^{3/4}]$, where $a^{\upharpoonleft} \in \mathbb{R}$ smoothly depends on ρ. Moreover, we see that, near ρz_j, the graph of the function

$$v^{\downharpoonleft} := u^{\downharpoonleft} + \tau X^{\downharpoonleft *}\Gamma_\rho$$

can be expanded as

$$v^{\downharpoonleft}(\rho z_j + y)$$
$$= -\sqrt{1-\rho^2} - \frac{\tau}{m+1}(m \log \rho + a^{\downharpoonleft}) - \frac{\tau}{m+1} \log |y| - \left(\rho - \frac{m}{m+1}\frac{\tau}{2\rho}\right) z_j \cdot y + \mathcal{O}_{\mathring{\mathscr{C}}^\infty}(\tau^{3/2})$$

for $|y| \in [(1/2)\tau^{3/4}, 2\tau^{3/4}]$, where $a^{\downharpoonleft} \in \mathbb{R}$ depends smoothly on ρ. The key point in our construction is that the constant in front of $z_j \cdot y$ can be adjusted by choosing ρ appropriately. Indeed, if we define $\rho_0 > 0$ by the identity

$$2(m+1)\rho_0^2 = m\tau,$$

then, when $\rho = \rho_0$, the constant in front of $z_j \cdot y$ in the last expansion is exactly 0, while choosing $\rho \neq \rho_0$ slightly larger or smaller allows one to prescribe any value of this constant, close enough to 0.

Mapping properties of the Jacobi operator about a punctured sphere. We first define on S^2 the distance function to the punctures $p^{\upharpoonleft}, p_0^{\downharpoonleft}, \dots, p_m^{\downharpoonleft}$ by

$$d := \operatorname{dist}_{S^2}(\,\cdot\,, \{p^{\upharpoonleft}, p_0^{\downharpoonleft}, \dots, p_m^{\downharpoonleft}\}).$$

Even though this is not apparent in the notation, the function d depends implicitly on ρ, since it depends on the location of the points $p_j^{\downharpoonleft}$, which themselves depend on ρ. We can define some weighted spaces on

$$S^* := S^2 - \{p^{\upharpoonleft}, p_0^{\downharpoonleft}, \dots, p_m^{\downharpoonleft}\}.$$

For $\nu \in \mathbb{R}$ and $k \in \mathbb{N}$, we define $\mathscr{C}_\nu^{k,\alpha}(S^*)$ to be the space of functions $w \in \mathscr{C}_{\mathrm{loc}}^{k,\alpha}(S^*)$ for which the following norm is finite:

$$\|w\|_{\mathscr{C}_\nu^{k,\alpha}(S^*)} := \sum_{j=0}^{k} \sup_{p \in S^*} d^{-\nu+j}(p) \|\nabla^j w(p)\|_{g_{S^2}} + \sup_{\zeta \in (0,\pi/2)} \sup_{d(p),d(q) \in [\zeta, 2\zeta]} \zeta^{-\nu+k+\alpha} \frac{\|\nabla^k w(p) - \nabla^k w(q)\|_{g_{S^2}}}{\operatorname{dist}_{S^2}(p,q)^\alpha}.$$

We further assume that the functions in $\mathscr{C}_\nu^{k,\alpha}(S^*)$ are invariant under the action of $\mathrm{Dih}_{m+1}^{(2)}$. Again, the weighted spaces $\mathscr{C}_\nu^{k,\alpha}(S^*)$ do implicitly depend on ρ.

We consider the operator

$$\mathbb{L}_\nu : \mathscr{C}^{2,\alpha}_\nu(S^*) \to \mathscr{C}^{0,\alpha}_{\nu-2}(S^*),$$
$$w \mapsto \mathbb{L}w.$$

It is easy to check that $\mathbb{L}_\nu$ is well defined.

Recall that $\mathbb{L}$ is conjugate to $\Delta_{S^2}+2$. When acting on a smooth function defined on S^2, the mapping properties of $\Delta_{S^2}+2$ are well understood, and we recall that the kernel of this operator is spanned by the restriction to S^2 of the linear forms on $\mathbb{R}^3$. Since we are assuming that the functions we consider are invariant under the action of the dihedral group $\mathrm{Dih}^{(2)}_{m+1}$, this implies that the bounded kernel of $\mathbb{L}$ has dimension 1. We now investigate the mapping properties of $\mathbb{L}$ (or, alternatively, $\Delta_{S^2}+2$) when acting on functions belonging to the weighted spaces we have just defined.

Proposition 6.5. *Assume that* $\nu \in (-1, 0)$. *Then there exist constants* $C, \rho_0 > 0$, *only depending on* ν, *such that, for all* $\rho \in (0, \rho_0)$, *we have*

$$\|w\|_{\mathscr{C}^{2,\alpha}_\nu(S^*)} \le C\|\mathbb{L}w\|_{\mathscr{C}^{0,\alpha}_\nu(S^*)},$$

for all functions w *in the* $L^2(S^2)$*-orthogonal complement of the functions* $\tilde{x}_3$ *and* Γ_ρ.

Proof. As usual, thanks to Schauder's estimates, it is enough to prove that

$$\|d^{-\nu}w\|_{L^\infty(S^*)} \le C\|d^{2-\nu}\mathbb{L}w\|_{L^\infty(S^*)}$$

for all ρ small enough.

As usual, the proof of this estimate is by contradiction. Assume that the estimate is not true. Then there exists a sequence ρ_n tending to 0 and a sequence of functions w_n such that

$$\|d^{-\nu}w_n\|_{L^\infty(S^*)} = 1 \quad \text{and} \quad \lim_{n\to\infty} \|d^{2-\nu}\mathbb{L}w_n\|_{L^\infty(S^*)} = 0.$$

Moreover w_n is invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$ and is L^2-orthogonal to $\tilde{x}_3$ and Γ_{ρ_n} (recall that $\Gamma_\rho = \Gamma_{\rho_n}$ depends on ρ_n). Hence

$$\int_{S^2} \tilde{x}_3 w_n \, d\mathrm{vol}_{S^2} = 0 \tag{6-5}$$

and

$$\int_{S^2} \Gamma_{\rho_n} w_n \, d\mathrm{vol}_{S^2} = 0. \tag{6-6}$$

We choose a point $q_n \in S^*$ such that

$$|w_n(q_n)| \ge \tfrac{1}{2} d^\nu(q_n),$$

and we distinguish various cases according to the behavior of the sequence $d(q_n)$. In each case, we rescale coordinates (using the exponential map) by $1/d(q_n)$, and we use elliptic estimates together with the Ascoli–Arzelà theorem to extract from the sequence $\tilde{w}_n := d^{-\nu}(q_n) w_n$ convergent subsequences. Finally, we pass to the limit in the equation satisfied by $\tilde{w}_n$. If, for some subsequence, $d(q_n)$ remains bounded away from 0, we get in the limit a nontrivial solution of

$$\mathbb{L} w = 0,$$

which is defined in $S^2 - \{p^{\upharpoonleft}, p^{\downharpoonleft}\}$, where we recall that $p^{\upharpoonleft}$ and $p^{\downharpoonleft}$ denote the north and south poles of S^2. Moreover, w is bounded by a constant times $(\mathrm{dist}(p, \{p^{\upharpoonleft}, p^{\downharpoonleft}\}))^\nu$ and w is invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$. Finally, we can pass to the limit in (6-5) and (6-6) and check that w is L^2-orthogonal to $\tilde{x}_3$ and $\Gamma_0 := \lim_{n\to\infty} \Gamma_{\rho_n}$. It is easy to check (see Remark 6.4) that this implies that $w \equiv 0$, which is a contradiction.

The second case we have to consider is the case where $\lim_{n\to\infty} d(q_n) = 0$ and $\lim_{n\to\infty} d(q_n)/\rho_n = +\infty$ or the case where $\lim_{n\to\infty} d(q_n)/\rho_n = 0$. In either case, we obtain a nontrivial solution of

$$\Delta w = 0$$

in $\mathbb{R}^2 - \{0\}$, which is bounded by a constant times $\mathrm{dist}(\,\cdot\,, \{0\})^\nu$. It is easy to check that $w \equiv 0$ since $\delta \notin \mathbb{Z}$, which is again a contradiction.

Finally, we consider the case where $\lim_{n\to\infty} d(q_n)/\rho_n$ exists. In this case, we obtain a nontrivial solution of $\Delta w = 0$ in $\mathbb{R}^2 - \{r_0 z_0, \dots, r_0 z_m\}$ for some $r_0 > 0$. Moreover, we know that this solution is bounded by a constant times $(\mathrm{dist}(\,\cdot\,, \{r_0 z_0, \dots, r_0 z_m\}))^\nu$ and w is also invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$. Inspection of the behavior of w at the points $r_0 z_j$ together with the fact that $\nu > -1$ and w is invariant with respect to the action of $\mathrm{Dih}^{(2)}_{m+1}$ implies that w is a solution in the sense of distributions of

$$\Delta w = a \sum_{j=0}^{m} \delta_{r_0 z_j}$$

for some $a \in \mathbb{R}$. Then inspection of w at infinity together with the fact that $\nu < 0$ implies that $a = 0$, and hence $w \equiv 0$. This is again a contradiction. □

Thanks to the previous result, we can prove:

Proposition 6.6. *Assume that $\nu \in (-1, 0)$ is fixed. Then the operator $\mathbb{L}_\nu$ is surjective and has a 2-dimensional kernel spanned by the functions $\tilde{x}_3$ and Γ_ρ. Moreover, the right inverse of $\mathbb{L}_\nu$, which is chosen so that its image is in the L^2-orthogonal complement of the kernel of $\mathbb{L}_\nu$, has a norm which is bounded independently of ρ small enough.*

Proof. The existence of a right inverse follows from the general theory developed, for example, in [Pacard 2008]. Nevertheless, we give a self-contained proof.

Let us assume that we are given a function $f \in \mathscr{C}^{0,\alpha}(S^*)$ which has compact support in S^*. Recall that the functions we are interested in are invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$. We choose $a \in \mathbb{R}$ so that $f - a\delta_{p^\upharpoonleft}$ is orthogonal to the function $\tilde{x}_3$. In particular, this implies that we can solve

$$\mathbb{L}\tilde{w} = f - a\delta_{p^\upharpoonleft},$$

and, choosing the constant $b \in \mathbb{R}$ appropriately, we can assume that $w := \tilde{w} - b\Gamma_\rho$ is L^2-orthogonal to the function $\tilde{x}_3$ and Γ_ρ. Observe that

$$\mathbb{L}w = f$$

in S^*, and also that $w \in \mathscr{C}^{2,\alpha}_\nu(S^*)$. In particular Proposition 6.5 applies and we have

$$\|w\|_{\mathscr{C}^{2,\alpha}_\nu(S^*)} \le C\|\mathbb{L}w\|_{\mathscr{C}^{0,\alpha}_\nu(S^*)}.$$

The general result, when f is not assumed to have compact support in S^*, can be handled as usual, using a sequence of functions having compact support and converging on compact sets to a given function in $\mathscr{C}^{0,\alpha}_\nu(S^*)$. □

A third fixed-point argument. Assume that we are given $\tau, \tilde{\tau} > 0$ small enough and satisfying

$$|\tilde{\tau} - \tau| \le \kappa\tau^{3/2}, \tag{6-7}$$

where the constant $\kappa > 0$ is large enough (it will be fixed in Section 7). We also assume that $\rho > 0$ satisfies

$$\left|\rho - \frac{m}{m+1}\frac{\tau}{2\rho}\right| \le \kappa\tau^{3/4}. \tag{6-8}$$

We prove the existence of an infinite-dimensional family of constant mean curvature surfaces which are close to $\mathfrak{S}_{\tau,\rho}$ and are parametrized by their boundary values described by two functions $f^\upharpoonleft : S^1 \to \mathbb{R}$ and $f^\downharpoonleft : S^1 \to \mathbb{R}$. The surfaces also depend on $\tilde{\tau}$ and ρ satisfying the above estimates.

Proposition 6.7. *Assume we are given $\kappa > 0$ large enough (to be fixed later). For all $\tau, \tilde{\tau} > 0$ small enough satisfying (6-7) and for all functions $f^\upharpoonleft$ which are invariant under the action of the dihedral group $\mathrm{Dih}^{(2)}_{m+1}$ and $f^\downharpoonleft$, which are invariant under the action of $\mathscr{S}_2$, satisfying both (H1) and*

$$\|f\|_{\mathscr{C}^{2,\alpha}(S^1)} \le \kappa\tau^{3/2},$$

there exists a constant mean curvature surface $\mathfrak{S}_{\tilde{\tau},\rho,f^\upharpoonleft,f^\downharpoonleft}$ which is a graph over $\mathfrak{S}_{\tau,\rho}$, has $m+2$ boundaries (one boundary close to the north pole and $m+1$

boundaries close to the south pole), and is invariant under the action of the dihedral group $\mathrm{Dih}^{(2)}_{m+1}$. *Close to the upper boundary, the surface* $\mathfrak{S}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}$ *is a* vertical graph *over the annulus*

$$\{x \in \mathbb{R}^2 : \tau^{3/4} \le |x| \le 2\tau^{3/4}\},$$

for some function $x \mapsto V^{\upharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}(\tau^{-3/4}x)$ *which can be expanded as*

$$\begin{aligned}(6\text{-}9)\quad & V^{\upharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}(x)\\ &= 1+\tilde\tau(m\log\rho + a^{\upharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}) + \tfrac{3}{4}\tilde\tau\log\tau + \tilde\tau\log|x| - W^{\text{out}}_{f^\upharpoonleft}(x) + \overline{V}^{\upharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}(x),\end{aligned}$$

where $a^\upharpoonleft \in \mathbb{R}$, W^{out}_f *denotes the bounded harmonic extension of* f *in* $\mathbb{R}^2 - \overline{D}(0,1)$, *and*

$$(6\text{-}10)\qquad \|\overline{V}^{\upharpoonleft}_{\tilde\tau,\rho,0,0}\|_{\mathscr{C}^{2,\alpha}(\overline{D}(0,2)-D(0,1))} \le C\tau^{3/2},$$

and, given $\nu \in (-1,0)$,

$$\begin{aligned}(6\text{-}11)\quad & \|\overline{V}^{\upharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft} - \hat{V}^{\upharpoonleft}_{\tilde\tau,\rho,f^{\upharpoonleft\prime},f^{\downharpoonleft\prime}}\|_{\mathscr{C}^{2,\alpha}\left(\overline{D}(0,1)-D\left(0,\frac{1}{2}\right)\right)}\\ &\qquad\le C\tau^{(1+\nu)/4}(\|f^{\upharpoonleft\prime}-f^\upharpoonleft\|_{\mathscr{C}^{2,\alpha}(S^1)} + \|f^{\downharpoonleft\prime}-f^\downharpoonleft\|_{\mathscr{C}^{2,\alpha}(S^1)})\end{aligned}$$

for some constant $C > 0$ *independent of* κ, $\tilde\tau$, *and* $f^\upharpoonleft$, $f^\downharpoonleft$, $f^{\upharpoonleft\prime}$, $f^{\downharpoonleft\prime}$.

Near one of the lower boundaries, the surface $\mathcal{S}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}$ *is a* vertical graph *over the annulus*

$$\{x \in \mathbb{R}^2 : \tau^{3/4} \le |x-\rho z_0| \le 2\tau^{3/4}\}$$

for some function $x \mapsto V^{\downharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}(\tau^{3/4}(x-\rho z_0))$, *which can be expanded as*

$$\begin{aligned}(6\text{-}12)\quad & V^{\downharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}(x)\\ &= -\sqrt{1-\rho^2} - \frac{\tilde\tau}{m+1}(m\log\rho + a^{\downharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}) - \frac{3\tilde\tau}{4(m+1)}\log\tau - \frac{\tilde\tau}{m+1}\log|x|\\ &\qquad - \tau^{3/4}\left(\rho - \frac{m}{m+1}\frac{\tilde\tau}{2\rho}\right)z_0\cdot x + W^{\text{out}}_{f^\downharpoonleft}(x) + \overline{V}^{\downharpoonleft}_{\tilde\tau,f^\upharpoonleft,f^\downharpoonleft}(x),\end{aligned}$$

where $\overline{V}^{\downharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}$ *enjoys properties similar to those described above for* $\overline{V}^{\downharpoonleft}_{\tilde\tau,\rho,f^\upharpoonleft,f^\downharpoonleft}$. *Moreover, both depend continuously on* $\tilde\tau$ *and* ρ.

Proof. Again the arguments of the proof are similar to those in the previous sections. The equation we must solve can be written formally as

$$(6\text{-}13)\qquad \mathbb{L}(\tilde\tau\Gamma_\rho + \hat{F} + w) = Q(\tilde\tau\Gamma_\rho + \hat{F} + w),$$

where Q collects all the nonlinear terms. Here $\hat{F}$ is a function which can be described as follows: near the north pole $p^{\upharpoonleft}$,

$$X^{\upharpoonleft *}\hat{F}(x) = \chi\, W^{\mathrm{out}}_{f^{\upharpoonleft}}(\tau^{-3/4}x),$$

where χ is a cutoff function identically equal to 1 in $D(0, 1/4)$ and identically equal to 1 outside $D\big(0, \frac{1}{2}\big)$. Near the south pole, $p^{\downharpoonleft}$

$$X^{\downharpoonleft *}\hat{F}(x) = \sum_{j=0}^{m} \bar{\chi}\left(\frac{x-\rho z_j}{\rho}\right) W^{\mathrm{out}}_{f^{\upharpoonleft}}(x-\rho z_j),$$

where $\bar{\chi}$ is a cutoff function identically equal to 1 in $D(0, c)$ and identically equal to 1 outside $D(0, c/2)$. Here $c = \sin(\pi/(m+1))$ so that the balls of radius c centered at the points z_j for $j = 0, \dots, m$ are disjoint.

We choose an extension operator

$$\hat{\mathscr{E}}_\tau : \mathscr{C}^{0,\alpha}(\mathfrak{S}_{\tau,\rho}) \to \mathscr{C}^{0,\alpha}(S^*)$$

such that

$$\hat{\mathscr{E}}_\tau(\psi) = \begin{cases} \psi & \text{in } \mathfrak{S}_{\tau,\rho}, \\ 0 & \text{in } X^{\upharpoonleft}(D(0, \tau^{3/4}/2)) \cup \bigcup_{j=0}^{m} X^{\downharpoonleft}(D(\rho z_j, \tau^{3/4}/2)), \end{cases}$$

and

$$\|\hat{\mathscr{E}}_\tau(\psi)\|_{\mathscr{C}^{0,\alpha}_\nu(S^*)} \le C\|\psi\|_{\mathscr{C}^{0,\alpha}_\nu(\mathfrak{S}_{\tau,\rho})}.$$

By definition, the norm in the space $\mathscr{C}^{0,\alpha}_\nu(\mathfrak{S}_{\tau,\rho})$ is defined exactly as the norm in $\mathscr{C}^{0,\alpha}_\nu(S^*)$, but points are restricted to $\mathfrak{S}_{\tau,\rho}$ instead of S^*.

We rewrite (6-13) as

$$\mathbb{L}w = \hat{\mathscr{E}}_\tau(-\mathbb{L}\hat{F} + Q(\tilde{\tau}\Gamma_\rho + \hat{F} + w)). \tag{6-14}$$

Observe that, by construction, $\mathbb{L}(\tilde{\tau}\Gamma_\rho) = 0$ away from the singular points.

Again, on the right side, it is understood that we consider the image by $\hat{\mathscr{E}}_\tau$ of the restriction of the functions to $\mathfrak{S}_{\tau,\rho}$.

We assume that $\nu \in (-1, 0)$ is fixed. It is easy to check that there exists a constant $c > 0$ (independent of κ) and a constant $c_\kappa > 0$ (depending on κ) such that

$$\|\hat{\mathscr{E}}_\tau(Q(\tilde{\tau}\Gamma_\rho))\|_{\mathscr{C}^{2,\alpha}_{\nu-2}(S^*)} \le c\,\tau^{(6-3\nu)/4},$$
$$\|\hat{\mathscr{E}}_\tau(\mathbb{L}\hat{F})\|_{\mathscr{C}^{2,\alpha}_{\nu-2}(S^*)} \le c\,\tau^{(1-2\nu)/4}(\|f^{\upharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)} + \|f^{\downharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)}),$$

and

$$\|\hat{\mathscr{E}}_\tau(Q(\tilde{\tau}\Gamma_\rho + \hat{F}' + w') - Q(\tilde{\tau}\Gamma_\rho + \hat{F} + w))\|_{\mathscr{C}^{2,\alpha}_{\nu-2}(S^*)}$$
$$\le c_\kappa(\tau\|w'-w\|_{\mathscr{C}^{2,\alpha}_\nu(S^*)} + \tau^{(4-3\nu)/4}(\|f^{\upharpoonleft\prime} - f^{\upharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)} + \|f^{\downharpoonleft\prime} - f^{\downharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)})),$$

provided w and w' satisfy

$$\|w\|_{\mathscr{C}^{2,\alpha}_\nu(S^*)} + \|w'\|_{\mathscr{C}^{2,\alpha}_\nu(S^*)} \le C\tau^{(6-3\nu)/4}$$

for some fixed constant $C > 0$ independent of κ. Here $\hat{F}$ and $\hat{F}'$ are associated to the harmonic extensions of the boundary data $f^{\restriction}$, $f^{\downharpoonleft}$ and $f^{\restriction\prime}$, $f^{\downharpoonleft\prime}$, respectively.

Now we make use of Proposition 6.6 to rephrase the problem as a fixed-point problem, and the previous estimates are precisely enough to solve this nonlinear problem using a fixed-point argument for contraction mappings in the ball of radius $C_\kappa\tau^{(6-3\nu)/4}$ in $\mathscr{C}^{2,\alpha}_\nu(S^*)$, where C_κ is fixed large enough. Then, for all $\tau > 0$ small enough, we find that there exists a constant $C_\kappa > 0$ (depending on κ) such that, for all functions $f^{\restriction}$, $f^{\downharpoonleft}$ satisfying the above hypothesis, there exists a solution w of (6-13) satisfying

$$\|w\|_{\mathscr{C}^{2,\alpha}_\nu(S^*)} \le C\tau^{(6-3\nu)/4}.$$

In addition, we have the estimate

$$\|w' - w\|_{\mathscr{C}^{2,\alpha}_\nu(S^*)} \le C_\kappa\tau^{(1-2\nu)/4}(\|f^{\restriction\prime} - f^{\restriction}\|_{\mathscr{C}^{2,\alpha}(S^1)} + \|f^{\downharpoonleft\prime} - f^{\downharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)})$$

for some constant $C > 0$, which does not depend on κ or τ, where w (respectively w') is the solution associated to $f^{\restriction}$, $f^{\downharpoonleft}$ (respectively $f^{\restriction\prime}$, $f^{\downharpoonleft\prime}$).

The solution of (6-13) is obtained through a fixed-point theorem for contraction mappings, and it is classical to check that the solution we obtain depends continuously on the parameters of the construction. In particular, the constant mean curvature surface we obtain depends continuously on $\tilde{\tau}$ and ρ. □

7. Connecting the pieces together

We keep the notation of the previous sections. We assume that $\kappa > 0$ is large enough (to be chosen shortly) and assume that $\tau > 0$ is chosen small enough so that all the results proven so far apply.

For all $\tilde{x} \in \mathbb{R}^2$, we define the annuli

$$A^{\mathrm{out}}_\tau(\tilde{x}) := \{x \in \mathbb{R}^2 : \tau^{3/4} \le |x - \tilde{x}| \le 2\tau^{3/4}\},$$
$$A^{\mathrm{ins}}_\tau(\tilde{x}) := \{x \in \mathbb{R}^2 : \tfrac{1}{2}\tau^{3/4} \le |x - \tilde{x}| \le \tau^{3/4}\}.$$

Recall from page 137 the conditions (H1) and (H2). Also recall that a function f defined on S^1 is invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$ if

$$f\left(\theta + \frac{2\pi}{m+1}\right) = f(\theta) \quad \text{for all } \theta \in S^1,$$

and f is invariant under the action of the symmetry $\mathscr{S}_2$ if

$$f(-\theta) = f(\theta) \quad \text{for all } \theta \in S^1.$$

We now describe the different pieces of constant mean curvature surfaces we have at hand.

(i) Assume that we are given $f^{\upharpoonleft} \in \mathscr{C}^{2,\alpha}(S^1)$, which is invariant under the action of $\mathrm{Dih}_{m+1}^{(2)}$ and satisfies (H1) and

$$\|f^{\upharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)} \leq \kappa \tau^{3/2}.$$

Proposition 4.12 provides a constant mean curvature (equal to 1) surface $\mathfrak{D}_{\tau,f}^{+}$ which is invariant under the action of the dihedral group $\mathrm{Dih}_{m+1}^{(2)}$, has one end asymptotic to the end of $\mathfrak{D}_{\tau}^{+}$, and which, close to its boundary, can be parametrized as the vertical graph of $x \mapsto U^{\upharpoonleft}(\tau^{-3/4}x)$ over $A_{\tau}^{\mathrm{ins}}(0)$, where

$$U^{\upharpoonleft}(x) = c^{\upharpoonleft} + \tau \log|x| - W_f^{\mathrm{ins}}(x) + \bar{U}^{\upharpoonleft}(x),$$

where

$$c^{\upharpoonleft} := \tau \log \frac{2}{\tau^{1/4}} \in \mathbb{R},$$

and where $\bar{U}^{\upharpoonleft}$ satisfies (4-17) and (4-18). To simplify the notation we have not mentioned the data τ, f in the notation for $U^{\upharpoonleft}$ and $\bar{U}^{\upharpoonleft}$.

(ii) Next, we assume that we are given $\tau_1 > 0$ satisfying

$$|\tau_1 - \tau| \leq \kappa \tau^{3/2},$$

and $\rho_1 > 0$ satisfying

$$\left| \rho_1 - \frac{m}{m+1} \frac{\tau}{2\rho_1} \right| \leq \kappa \tau^{3/4}.$$

Further assume that we are given a function $f_1^{\upharpoonleft} \in \mathscr{C}^{2,\alpha}(S^1)$ invariant under the action of the dihedral group $\mathrm{Dih}_{m+1}^{(2)}$ and a function $f_1^{\downharpoonleft} \in \mathscr{C}^{2,\alpha}(S^1)$ invariant under the action of the symmetry $\mathscr{S}_2$, both satisfying (H1) and

$$\|f_1^{\upharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)} \leq \kappa \tau^{3/2} \quad \text{and} \quad \|f_1^{\downharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)} \leq \kappa \tau^{3/2}.$$

Proposition 6.7 provides a constant mean curvature (equal to 1) surface $\mathfrak{S}_{\tau_1,\rho_1,f_1^{\upharpoonleft},f_1^{\downharpoonleft}}$ which is invariant under the action of the dihedral group $\mathrm{Dih}_{m+1}^{(2)}$ and which, close to its upper boundary, can be parametrized as the vertical graph of $x \mapsto V^{\upharpoonleft}(\tau^{-3/4}x)$ over $A_{\tau}^{\mathrm{out}}(0)$, where

$$V^{\upharpoonleft}(x) = 1 + d^{\upharpoonleft} + \tau_1 \log|x| - W_{f_1^{\upharpoonleft}}^{\mathrm{out}}(x) + \bar{V}^{\upharpoonleft}(x),$$
$$d^{\upharpoonleft} := \tau_1 (m \log \rho_1 + a_{\tau_1,\rho,f_1^{\upharpoonleft},f_1^{\downharpoonleft}}^{\upharpoonleft} + \tfrac{3}{4} \log \tau) \in \mathbb{R},$$

and where $\bar{V}^{\upharpoonleft}$ satisfies (6-10) and (6-11). Close to one of its lower boundaries, this surface can be parametrized as a vertical graph for some function

$$x \mapsto V^{\downharpoonleft}(\tau^{-3/4}(x - \rho_1 z_0))$$

over $A_\tau^{\text{out}}(\rho_1 z_0)$ which can be expanded as

$$V^{\downharpoonleft}(x) = -1 + c^{\downharpoonleft} - \frac{\tau_1}{m+1}\log|x| - \tau^{3/4}\left(\rho_1 - \frac{m}{m+1}\frac{\tau_1}{2\rho}\right) z_0 \cdot x + W^{\text{out}}_{f_1^{\downharpoonleft}}(x) + \bar{V}^{\downharpoonleft}(x),$$
$$c^{\downharpoonleft} := 1 - \sqrt{1-\rho_1^2} - \frac{\tau_1}{m+1}(m\log\rho_1 + a^{\downharpoonleft}_{\tau_1,\rho_1,f_1^{\upharpoonleft},f_1^{\downharpoonleft}} - \tfrac{3}{4}\log\tau) \in \mathbb{R},$$

and where $\bar{V}^{\downharpoonleft}$ satisfies estimates of the form (6-10) and (6-11). Again, to simplify the notation we have not mentioned the parameters τ_1, ρ_1, $f_1^{\upharpoonleft}$, $f_1^{\downharpoonleft}$ in the notation for $V^{\upharpoonleft}$, $\bar{V}^{\upharpoonleft}$, $V^{\downharpoonleft}$, and $\bar{V}^{\downharpoonleft}$.

(iii) Assume that we are given $\tau_2 > 0$ satisfying

$$\left|\tau_2 - \frac{\tau}{m+1}\right| \le \kappa\tau^{3/2},$$

and a function $f_2^{\downharpoonleft} \in \mathscr{C}^{2,\alpha}(S^1)$ which satisfies (H1), (H2), and

$$\|f_2^{\downharpoonleft}\|_{\mathscr{C}^{2,\alpha}(S^1)} \le \kappa\tau^{3/2}.$$

Proposition 5.5 provides a constant mean curvature (equal to 1) surface $\mathfrak{C}_{\tau_2, f_2^{\downharpoonleft}}$ which is invariant under the action of $\mathscr{S}_3$, the symmetry with respect to the horizontal plane $x_3 = 0$, and is also invariant under the action of $\mathscr{S}_2$, the symmetry with respect to the plane $x_2 = 0$. Moreover, close to its lower boundary, this surface can be parametrized as the vertical graph of $x \mapsto U^{\downharpoonleft}(\tau^{-3/4}x)$ over $A_\tau^{\text{ins}}(0)$, where

$$U^{\downharpoonleft}(x) = d^{\downharpoonleft} - \tau_2 \log|x| + W^{\text{ins}}_{f_2^{\downharpoonleft}}(x) + \bar{U}^{\downharpoonleft}(x),$$

where

$$d^{\downharpoonleft} := -\tau_2 \log\frac{2\tau^{3/4}}{\tau_2} \in \mathbb{R},$$

and where $\bar{U}^{\downharpoonleft}$ satisfies (5-8) and (5-9). To simplify the notation we have not mentioned the data τ_2, $f_2^{\downharpoonleft}$ in the notation for $\bar{U}^{\downharpoonleft}$ and $\bar{U}^{\downharpoonleft}$.

Let us emphasize that the functions $f_1^{\downharpoonleft}$, $f_2^{\downharpoonleft}$ and $f^{\upharpoonleft}$, $f_1^{\upharpoonleft}$ are all assumed to satisfy (H1). Hence they have no constant term in their Fourier series. The function $f_2^{\downharpoonleft}$ is also assumed to satisfy (H2). Now, the functions $f^{\upharpoonleft}$ and $f_1^{\upharpoonleft}$ are assumed to be invariant under the action of the dihedral group $\mathrm{Dih}^{(2)}_{m+1}$, and, as was already mentioned, this implies that both functions also satisfies (H2) since its Fourier series does not contain any term of the form $z \cdot x$. Therefore, $f_1^{\downharpoonleft}$ is the only function which does not satisfy (H2). Since $f_1^{\downharpoonleft}$ is assumed to be invariant under the action of $\mathscr{S}_2$, we can decompose it as

$$f_1^{\downharpoonleft} = \lambda_1 z_0 \cdot x + f_1^{\downharpoonleft,\perp},$$

where $\lambda_1 \in \mathbb{R}$ and where $f_1^{\downharpoonleft,\perp}$ satisfies both (H1) and (H2).

We denote by $\mathfrak{C}^{(0)}_{\tau_2, f_2^{\downharpoonleft}, \rho_1}$ the surface $\mathfrak{C}_{\tau_2, f_2^{\downharpoonleft}}$ translated by $\rho_1 z_0$. For $j = 1, \dots, m$,

$$\mathfrak{C}^{(0)}_{\tau_2, f_2^\downharpoonleft, \rho_1} := \mathfrak{C}_{\tau_2, f_2^\downharpoonleft} + \rho_1 z_0.$$

We denote the image of $\mathfrak{C}^{(0)}_{\tau_2, f_2^\downharpoonleft, \rho_1}$ under the rotation $(\mathcal{R}_{m+1})^j$ by

$$\mathfrak{C}^{(j)}_{\tau_2, f_2^\downharpoonleft, \rho_1} := (\mathcal{R}_{m+1})^j (\mathfrak{C}^{(0)}_{\tau_2, f_2^\downharpoonleft, \rho_1}).$$

In particular, the collection of surfaces $\mathfrak{C}^{(0)}_{\tau_2, f_2^\downharpoonleft, \rho_1}, \dots, \mathfrak{C}^{(m)}_{\tau_2, f_2^\downharpoonleft, \rho_1}$ constitute $m+1$ constant mean curvature surfaces which are symmetric with respect to the dihedral group $\mathrm{Dih}^{(3)}_{m+1}$.

Given $t_1 \in \mathbb{R}$ small enough, we denote by $\mathfrak{S}_{\tau_1, \rho_1, f_1^\upharpoonleft, f_1^\downharpoonleft, t_1}$ the surface $\mathfrak{S}_{\tau_1, \rho_1, f_1^\upharpoonleft, f_1^\downharpoonleft}$ translated in the vertical direction by $(1 - c^\downharpoonleft + d^\downharpoonleft + t_1) e_3$:

$$\mathfrak{S}_{\tau_1, \rho_1, f_1^\upharpoonleft, f_1^\downharpoonleft, t_1} := \mathfrak{S}_{\tau_1, \rho_1, f_1^\upharpoonleft, f_1^\downharpoonleft} + (1 - c^\downharpoonleft + d^\downharpoonleft + t_1) e_3.$$

This is a constant mean curvature surface which is symmetric with respect to the dihedral group $\mathrm{Dih}^{(2)}_{m+1}$. Observe that the lower boundaries of $\mathfrak{C}^{(0)}_{\tau_2, f_2^\downharpoonleft, \rho_1}, \dots, \mathfrak{C}^{(m)}_{\tau_2, f_2^\downharpoonleft, \rho_1}$ are close to the lower boundaries of $\mathfrak{S}_{\tau_1, \rho_1, f_1^\upharpoonleft, f_1^\downharpoonleft, t_1}$.

Finally, given $t \in \mathbb{R}$ small enough, we denote by $\mathfrak{D}^+_{\tau, f, t}$ the surface $\mathfrak{D}^+_{\tau, f, t}$ translated in the vertical direction by $(2 - c^\downharpoonleft + d^\downharpoonleft - c^\upharpoonleft + d^\upharpoonleft + t_1 + t) e_3$:

$$\mathfrak{D}^+_{\tau, f, t} := \mathfrak{D}^+_{\tau, f} + (2 - c^\downharpoonleft + d^\downharpoonleft - c^\upharpoonleft + d^\upharpoonleft + t_1 + t) e_3.$$

This is a constant mean curvature surface which is symmetric with respect to the dihedral group $\mathrm{Dih}^{(2)}_{m+1}$. Observe that the boundary of $\mathfrak{D}^+_{\tau, f, t}$ is close to the upper boundary of $\mathfrak{S}_{\tau_1, \rho_1, f_1^\upharpoonleft, f_1^\downharpoonleft, t_1}$.

To complete the proof of the main theorem, it remains to adjust the free parameters of our construction, namely, $t, t_1, \tau_1, \tau_2, \rho_1 \in \mathbb{R}$, and the functions $f_1^\downharpoonleft$, $f_2^\downharpoonleft$, $f^\upharpoonleft$ and $f_1^\upharpoonleft$ defined on S^1, so that

$$\mathfrak{C}^{(0)}_{\tau_2, f_2^\downharpoonleft, \rho_1} \sqcup \cdots \sqcup \mathfrak{C}^{(m)}_{\tau_2, f_2^\downharpoonleft, \rho_1} \sqcup \mathfrak{S}_{\tau_1, \rho_1, f_1^\upharpoonleft, f_1^\downharpoonleft, t_1} \sqcup \mathfrak{D}^+_{\tau, f, t}$$

constitute a $\mathcal{C}^1$ surface which can be extended by reflection through the horizontal plane as a $\mathcal{C}^1$ surface which is complete, noncompact, and has two ends of Delaunay type (asymptotic to a nodoid end). Observe that the surface is invariant under the action of the dihedral group $\mathrm{Dih}^{(3)})_{m+1}$ and that there is still one free parameter, namely, τ, which determines the Delaunay type end and hence the vertical flux of the surface.

This surface is in fact piecewise smooth and has constant mean curvature equal to 1 away from the boundaries where the connected sum is performed. Since all pieces have constant mean curvature identically equal to 1, elliptic regularity theory then implies that this surface is in fact a smooth surface. Indeed, near one of the boundaries where the connected sum is performed the surface is a graph of a function, say u^{ins} defined over A_τ^{ins} and another function, say u^{out} defined over A_τ^{out}. The functions u^{ins} and u^{out} are $\mathscr{C}^{2,\alpha}$ and solve the mean curvature equation

$$\tfrac{1}{2}\operatorname{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right)=1 \tag{7-1}$$

on their respective domains of definition (for the sake of simplicity, we assume that the mean curvature vector is upward pointing near the boundary we are interested in). Moreover, $u^{\text{ins}}=u^{\text{out}}$ and $\partial_r u^{\text{ins}}=\partial_r u^{\text{out}}$ on $\partial A_\tau^{\text{ins}}\cap\partial A_\tau^{\text{out}}$. This implies that the function u defined on $A_\tau^{\text{ins}}\cup A_\tau^{\text{out}}$ by $u:=u^{\text{ins}}$ on A_τ^{ins} and $u:=u^{\text{out}}$ on A_τ^{out} belongs to $\mathscr{C}^{1,1}$ and is a weak solution of (7-1) on $A_\tau^{\text{ins}}\cup A_\tau^{\text{out}}$. Elliptic regularity implies that u is $\mathscr{C}^{2,\alpha}$, and hence the surface we have obtained is a smooth constant mean curvature surface.

Therefore, to complete the proof, it remains to explain how to find the parameters $t, t_1, \tau_1, \tau_2, \rho_1\in\mathbb{R}$, and the functions $f_1^{\downharpoonleft}$, $f_2^{\downharpoonleft}$, $f^{\upharpoonleft}$, and $f_1^{\upharpoonleft}$ defined on S^1, so that the following system of equations on S^1 is fulfilled:

$$U^{\upharpoonleft}-c^{\upharpoonleft}+t=V^{\upharpoonleft}-1-d^{\upharpoonleft}\quad\text{and}\quad\partial_r(V^{\upharpoonleft}-U^{\upharpoonleft})=0, \tag{7-2}$$

$$V^{\downharpoonleft}+1-c^{\downharpoonleft}+t_1=U^{\downharpoonleft}-d^{\downharpoonleft}\quad\text{and}\quad\partial_r(V^{\downharpoonleft}-U^{\downharpoonleft})=0. \tag{7-3}$$

Recall that, even though this is not apparent in the notations, all functions and constants depend on the parameters and boundary data. The rest of this section is devoted to the proof that the above system does have a solution, provided τ is small enough.

Proposition 7.1. *There exists $\kappa>0$ such that, for all $\tau>0$ small enough, there exist parameters $t, t_1, \tau_1, \tau_2, \rho_1$, and functions $f_1^{\downharpoonleft}$, $f_2^{\downharpoonleft}$, $f^{\upharpoonleft}$, $f_1^{\upharpoonleft}$ defined on S^1 and satisfying the above symmetries and estimates such that the system (7-2) and (7-3) is satisfied.*

Proof. First we make use of Propositions 4.12, 5.5, and 6.7 to get the expansion of the functions $U^{\upharpoonleft}$, $V^{\upharpoonleft}$, $U^{\downharpoonleft}$ and $V^{\downharpoonleft}$. Recalling that we have to restrict all those functions to S^1, it is easy to check, using (4-16) and (6-9), that the first two equations of the system we have to solve read

$$\begin{aligned}&t+f_1^{\upharpoonleft}-f^{\upharpoonleft}=\overline{V}^{\upharpoonleft}-\overline{U}^{\upharpoonleft},\\&(\tau_1-\tau)+\partial_r(W^{\text{out}}_{f_1^{\upharpoonleft}}-W^{\text{ins}}_{f^{\upharpoonleft}})=\partial_r(\overline{V}^{\upharpoonleft}-\overline{U}^{\upharpoonleft}),\end{aligned} \tag{7-4}$$

while, using (5-7) and (6-12), we see that the next two equations are given by

$$\begin{aligned}&t_1-\tau^{3/4}\left(\rho_1-\frac{m}{m+1}\frac{\tau_1}{2\rho_1}\right)z_0\cdot x+f_1^{\downharpoonleft}-f_2^{\downharpoonleft}=\overline{U}^{\downharpoonleft}-\overline{V}^{\downharpoonleft},\\&-\left(\frac{\tau_1}{m+1}-\tau_2\right)-\tau^{3/4}\left(\rho_1-\frac{m}{m+1}\frac{\tau_1}{2\rho_1}\right)z_0\cdot x+\partial_r(W^{\text{out}}_{f_1^{\downharpoonleft}}-W^{\text{ins}}_{f_2^{\downharpoonleft}})\\&\qquad\qquad=\partial_r(\overline{U}^{\downharpoonleft}-\overline{V}^{\downharpoonleft}).\end{aligned} \tag{7-5}$$

In writing this system one has to be a bit careful about the invariance of the functions we are interested in. Indeed, in (7-4), all functions are invariant under the action of $\mathrm{Dih}^{(2)}_{m+1}$, while in (7-5), all functions are invariant under the action of the symmetry $\mathcal{S}_2$.

Let us denote by Π^0 the $L^2(S^1)$-orthogonal projection over the space of constant functions, by Π^1, the $L^2(S^1)$-orthogonal projection over the space spanned by the function $x \mapsto z_0 \cdot x$, and by $\Pi^\perp$, the $L^2(S^1)$-orthogonal projection over the orthogonal complement of the space spanned by the constant function and the function $x \mapsto z_0 \cdot x$.

We project this system over the $L^2(S^1)$-orthogonal complement of the constant function and the function $x \mapsto z_0 \cdot x$. We obtain the coupled system

$$
\begin{aligned}
& f_1^{\upharpoonleft} - f^{\upharpoonleft} = \Pi^\perp(\overline{V}^{\upharpoonleft} - \overline{U}^{\upharpoonleft}),\\
& \partial_r(W^{\mathrm{out}}_{f_1^{\upharpoonleft}} - W^{\mathrm{ins}}_{f^{\upharpoonleft}}) = \Pi^\perp \partial_r(\overline{V}^{\upharpoonleft} - \overline{U}^{\upharpoonleft}),\\
& f_1^{\downharpoonleft,\perp} - f_2^{\downharpoonleft} = \Pi^\perp(\overline{U}^{\downharpoonleft} - \overline{V}^{\downharpoonleft}),\\
& \partial_r(W^{\mathrm{out}}_{f_1^{\downharpoonleft,\perp}} - W^{\mathrm{ins}}_{f_2^{\downharpoonleft}}) = \Pi^\perp \partial_r(\overline{U}^{\downharpoonleft} - \overline{V}^{\downharpoonleft}),
\end{aligned}
\tag{7-6}
$$

where we recall that we have decomposed $f_1^{\downharpoonleft} = \lambda_1 z_0 \cdot x + f_1^{\downharpoonleft,\perp}$.

The projection of the system (7-4)–(7-5) over the space of constant functions leads to the coupled system

$$
\begin{aligned}
& t = \Pi^0(\overline{V}^{\upharpoonleft} - \overline{U}^{\upharpoonleft}),\\
& \tau_1 - \tau = \Pi^0 \partial_r(\overline{V}^{\upharpoonleft} - \overline{U}^{\upharpoonleft}),\\
& t_1 = \Pi^0(\overline{U}^{\downharpoonleft} - \overline{V}^{\downharpoonleft}),\\
& \tau_2 - \frac{\tau_1}{m+1} = \Pi^0 \partial_r(\overline{U}^{\downharpoonleft} - \overline{V}^{\downharpoonleft}).
\end{aligned}
\tag{7-7}
$$

Finally, the projection of the system (7-4)–(7-5) over the space of functions spanned by $x \mapsto z_0 \cdot x$ leads to the coupled system

$$
\begin{aligned}
& \left(\lambda_1 - \tau^{3/4}\left(\rho_1 - \frac{m}{m+1}\frac{\tau_1}{2\rho_1}\right)\right) z_0 \cdot x = \Pi^1(\overline{U}^{\downharpoonleft} - \overline{V}^{\downharpoonleft}),\\
& \left(-\lambda_1 - \tau^{3/4}\left(\rho_1 - \frac{m}{m+1}\frac{\tau_1}{2\rho_1}\right)\right) z_0 \cdot x = \Pi^1 \partial_r(\overline{U}^{\downharpoonleft} - \overline{V}^{\downharpoonleft}).
\end{aligned}
\tag{7-8}
$$

To obtain the second equation, we have used the fact that

$$
W^{\mathrm{out}}_{f_1^{\downharpoonleft}} = \lambda_1 \frac{z_0 \cdot x}{|x|^2} + W^{\mathrm{out}}_{f_1^{\downharpoonleft,\perp}}.
$$

Observe that the right sides of (7-6), (7-7), and (7-8) do not depend on t and t_1. Hence the first and third equations in (7-7) give us the values of t and t_1, once the rest of the equations are solved.

For all τ small enough, we solve (7-6) using a fixed-point theorem for contraction mappings to obtain a solution $(f^{\upharpoonleft}, f_1^{\upharpoonleft}, f_1^{\downharpoonleft,\perp}, f_2^{\downharpoonleft})$ continuously depending on the parameters τ_1, τ_2, ρ_1, λ_1 (and τ). Then we introduce the corresponding solution in (7-7) and (7-8) to get a nonlinear system in τ_1, τ_2, and ρ_1, which we solve using Browder's fixed-point theorem.

To begin with, we explain how (7-6) can be rewritten in diagonal form. This makes use of the following result, whose proof can be found, for example, in [Mazzeo and Pacard 2001]:

Proposition 7.2. *The operator*

$$\mathscr{C}^{2,\alpha}(S^1)^\perp \ni f \mapsto \partial_r(W_f^{\text{ins}} - W_f^{\text{out}})_{|r=1} \in \mathscr{C}^{1,\alpha}(S^1)^\perp$$

is an isomorphism. Here $\mathscr{C}^{k,\alpha}(S^1)^\perp$ *denotes the image of* $\mathscr{C}^{k,\alpha}(S^1)$ *under* $\Pi^\perp$.

Proof. The Fourier decomposition of a function $f \in \mathscr{C}^{k,\alpha}(S^1)^\perp$ is given by

$$f(\theta) = \sum_{n\neq 0,\pm 1} f_n e^{in\theta},$$

in which case

$$W_f^{\text{out}} = \sum_{n\neq 0,\pm 1} f_n r^{-|n|} e^{in\theta} \quad \text{and} \quad W_f^{\text{ins}} = \sum_{n\neq 0,\pm 1} f_n r^{|n|} e^{in\theta}.$$

Therefore,

$$\partial_r(W_f^{\text{ins}} - W_f^{\text{out}})_{|r=1} = 2 \sum_{n\neq 0,\pm 1} f_n |n| e^{in\theta}$$

is equal to twice the Dirichlet to Neumann map for the Laplace operator in the unit disc. This is a well-defined, selfadjoint, first order elliptic operator which is injective, and elliptic regularity theory implies that it is an isomorphism. □

Using this result, the system (7-6) can be rewritten as

$$(f^{\upharpoonleft}, f_1^{\upharpoonleft}, f_1^{\downharpoonleft,\perp}, f_2^{\downharpoonleft}) = \mathbb{N}^\perp_{\tau_1,\tau_2,\rho_1,\lambda_1}(f^{\upharpoonleft}, f_1^{\upharpoonleft} f_1^{\downharpoonleft,\perp}, f_2^{\downharpoonleft}),$$

where the nonlinear operator $\mathbb{N}^\perp_{\tau_1,\tau_2,\rho_1,\lambda_1}$ satisfies

$$\|\mathbb{N}^\perp_{\tau_1,\tau_2,\rho_1,\lambda_1}(f^{\upharpoonleft}, f_1^{\upharpoonleft}, f_1^{\downharpoonleft,\perp}, f_2^{\downharpoonleft})\|_{(\mathscr{C}^{2,\alpha}(S^1))^4} \leq C\tau^{3/2} \tag{7-9}$$

for some constant $C > 0$ independent of $\kappa > 0$, provided τ is chosen small enough. This last estimate follows directly from (4-17), (5-8) and (6-10). Moreover, thanks to (4-18), (5-9), and (6-11), provided $\kappa > 0$ is fixed larger than the constant C which appears in (7-9), we can use a fixed-point theorem for contraction mappings in the ball of radius $\kappa\tau^{3/2}$ in $\left(\Pi^\perp\mathscr{C}^{2,\alpha}(S^1)\right)^4$ to get the existence of a solution of (7-9) for all $\tau > 0$ small enough. This solution depends continuously on τ_1, τ_2, ρ_1, and λ_1, since $\mathbb{N}^\perp_{\tau_1,\tau_2,\rho_1,\lambda_1}$ does (observe that $\mathbb{N}^\perp_{\tau_1,\tau_2,\rho_1,\lambda_1}$ depends implicitly on

τ). We now insert this solution in (7-7) and (7-8). With simple manipulations, we conclude that it remains to solve the nonlinear system

$$\left(\tau_1-\tau,\, \tau_2-\frac{\tau}{m+1},\, \tau^{3/4}\left(\rho_1-\frac{m}{m+1}\frac{\tau}{2\rho_1}\right),\, \lambda_1\right)=\mathbb{N}^0(\tau_1,\tau_2,\rho_1,\lambda_1), \tag{7-10}$$

where $\mathbb{N}^0$ satisfies

$$\|\mathbb{N}^0(\tau_1,\tau_2,\rho_1,\lambda_1)\|_{\mathbb{R}^4}\le C\tau^{3/2}$$

for some constant $C>0$ independent of $\kappa>0$, provided τ is chosen small enough. Moreover, $\mathbb{N}^0$ depends continuously on the parameters τ_1, τ_2, ρ_1, and λ_1 (observe that $\mathbb{N}^0$ depends implicitly on τ). Equation (7-10) can then be solved using a simple degree argument (Browder's fixed-point theorem). $\square$

Appendix

We discuss the elementary result in the theory of second order ordinary differential equations which is used at the end of the proof of Proposition 4.8. Assume that we are given a function $s\mapsto p(s)$ which is periodic (say of period $S>0$). Further assume that the homogeneous problem $(\partial_s^2+p)w^+=0$ has a nontrivial periodic solution of period S. Without loss of generality, we can assume that $w^+(0)=1$ and $\partial_s w^+(0)=0$ (just choose the origin so that 0 coincides with a point where w^+ achieves its maximum). Let w^- be the unique solution of $(\partial_s^2+p)w^-=0$ such that $w^-(0)=0$ and $\partial_s w^-(0)=1$. The Wronskian of w^+ and w^- being constant, we conclude that

$$\begin{aligned}\partial_s w^-(S)&=\partial_s w^-(S)w^+(S)-\partial_s w^+(S)w^-(S)\\&=\partial_s w^-(0)w^+(0)-\partial_s w^+(0)w^-(0)\\&=1.\end{aligned}$$

We define

$$v(s):=w^-(S+s)-w^-(S)w^+(s).$$

It is clear that v is a solution of $(\partial_s^2+p)v=0$. Further, observe that $\partial_s v(0)=1$ and $v(0)=0$. Therefore, $v=w^-$. This proves that

$$w^-(S+s)=w^-(s)+w^-(S)w^+(s),$$

and hence w^- is at most growing linearly in the sense that $|w^-(s)|\le C(1+|s|)$ for some constant $C>0$.

References

[Colding and Minicozzi 2011] T. H. Colding and W. P. Minicozzi, II, *A course in minimal surfaces*, Graduate Studies in Mathematics **121**, American Mathematical Society, Providence, RI, 2011. MR 2780140 Zbl 1242.53007

[Eells 1987] J. Eells, "The surfaces of Delaunay", *Math. Intelligencer* **9**:1 (1987), 53–57. MR 88h:53011 Zbl 0605.53002

[Fakhi and Pacard 2000] S. Fakhi and F. Pacard, "Existence result for minimal hypersurfaces with a prescribed finite number of planar ends", *Manuscripta Math.* **103**:4 (2000), 465–512. MR 2002d:53017 Zbl 0992.53011

[Hauswirth and Pacard 2007] L. Hauswirth and F. Pacard, "Higher genus Riemann minimal surfaces", *Invent. Math.* **169**:3 (2007), 569–620. MR 2008i:58013 Zbl 1129.53009

[Lawson 1977] H. B. Lawson, Jr., *Lectures on minimal submanifolds, I*, Monografías de Matemática **14**, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1977. MR 82d:53035a Zbl 0924.60012

[Lockhart and McOwen 1985] R. B. Lockhart and R. C. McOwen, "Elliptic differential operators on noncompact manifolds", *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **12**:3 (1985), 409–447. MR 87k:58266 Zbl 0615.58048

[Mazzeo 1991] R. Mazzeo, "Elliptic theory of differential edge operators, I", *Comm. Partial Differential Equations* **16**:10 (1991), 1615–1664. MR 93d:58152 Zbl 0745.58045

[Mazzeo and Pacard 2001] R. Mazzeo and F. Pacard, "Constant mean curvature surfaces with Delaunay ends", *Comm. Anal. Geom.* **9**:1 (2001), 169–237. MR 2002b:53005 Zbl 1005.53006

[Mazzeo et al. 2001] R. Mazzeo, F. Pacard, and D. Pollack, "Connected sums of constant mean curvature surfaces in Euclidean 3 space", *J. Reine Angew. Math.* **536** (2001), 115–165. MR 2002d:53020 Zbl 0972.53010

[Mazzeo et al. 2005] R. Mazzeo, F. Pacard, and D. Pollack, "The conformal theory of Alexandrov embedded constant mean curvature surfaces in $\mathbb{R}^3$", pp. 525–559 in *Global theory of minimal surfaces*, edited by D. Hoffman, Clay Math. Proc. **2**, Amer. Math. Soc., Providence, RI, 2005. MR 2006g:53009 Zbl 1101.53006

[Melrose 1993] R. B. Melrose, *The Atiyah-Patodi–Singer index theorem*, Research Notes in Mathematics **4**, A K Peters Ltd., Wellesley, MA, 1993. MR 96g:58180 Zbl 0796.58050

[Pacard 2008] F. Pacard, "Connected sum constructions in geometry and nonlinear analysis", Unpublished lecture notes, 2008, Available at http://www.math.polytechnique.fr/~pacard/Publications/Lecture-Part-I.pdf.

[Rossman 2005] W. Rossman, "The first bifurcation point for Delaunay nodoids", *Experiment. Math.* **14**:3 (2005), 331–342. MR 2006j:53011 Zbl 1092.53010

Received July 31, 2012. Revised December 16, 2012.

FRANK PACARD
CENTRE DE MATHÉMATIQUES LAURENT SCHWARTZ, UMR-CNRS 7640
ÉCOLE POLYTECHNIQUE
PALAISEAU, 91128
FRANCE

frank.pacard@math.polytechnique.fr

HAROLD ROSENBERG
INSTITUTO DE MATEMÁTICA PURA E APLICADA
110 ESTRADA DONA CASTORINA
RIO DE JANEIRO 22460-320
BRAZIL

hrosen@free.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.185

SOME NEW CANONICAL FORMS FOR POLYNOMIALS

BRUCE REZNICK

We give some new canonical representations for forms over $\mathbb{C}$. For example, a general binary quartic form can be written as the square of a quadratic form plus the fourth power of a linear form. A general cubic form in $(x_1, \dots, x_n)$ can be written uniquely as a sum of the cubes of linear forms $l_{ij}(x_i, \dots, x_j)$, $1 \le i \le j \le n$. A general ternary quartic form is the sum of the square of a quadratic form and three fourth powers of linear forms. The methods are classical and elementary.

1. Introduction and overview

Introduction. Let $H_d(\mathbb{C}^n)$ denote the vector space of complex forms of degree d in n variables, or *n-ary d-ic forms*; it has dimension $N(n,d) := \binom{n+d-1}{d}$. One of the major accomplishments of 19th century algebra was the discovery of canonical forms for certain classes of n-ary d-ics, especially as the sum of d-th powers of linear forms. By a *canonical form* we mean a polynomial $F(t; x)$ in two sets of variables, $t \in \mathbb{C}^{N(n,d)}$ and $x \in \mathbb{C}^n$, with the property that, for general $p \in H_d(\mathbb{C}^n)$, there exists t so that $p(x) = F(t; x)$. Put another way, the set $\{F(t; x) : t \in \mathbb{C}^{N(n,d)}\}$ is a Zariski open set in $H_d(\mathbb{C}^n)$.

In this paper, we present some new canonical forms, whose main novelty is that they involve intermediate powers of forms of higher degree, or forms with a restricted set of monomials. (These variations have been suggested by Hilbert's study of ternary quartics [1888], which led to his 17th problem, as well as by a remarkable theorem of B. Reichstein [1987] on cubic forms.) These expressions are less susceptible to apolarity arguments than the traditional canonical forms, and lead naturally to (mostly open) enumeration questions.

To take a simple, yet familiar, example,

$$F(t_1, t_2, t_3; x, y) = (t_1 x + t_2 y)^2 + (t_3 y)^2 \tag{1-1}$$

is a canonical form for binary quadratic forms. By the usual completion of squares, $p(x, y) = ax^2 + 2bxy + cy^2$ can be put into (1-1) for $t_1 = \sqrt{a}$, $t_2 = b/t_1$ and

MSC2000: primary 11E76, 14N15; secondary 11E25, 11P05, 15A72.
Keywords: canonical forms, sums of powers of polynomials.

$t_3^2 = c - t_2^2$. Many of the examples in this paper can be viewed as imperfect attempts to generalize (1-1).

Sylvester [1851a; 1851b] presented a family of canonical forms for binary forms in all degrees.

Theorem 1.1 (Sylvester's theorem). (i) *A general binary form p of odd degree $2s-1$ can be written as*

$$p(x, y) = \sum_{j=1}^{s} (\alpha_j x + \beta_j y)^{2s-1}. \tag{1-2}$$

(ii) *A general binary form p of even degree $2s$ can be written as*

$$p(x, y) = \lambda x^{2s} + \sum_{j=1}^{s} (\alpha_j x + \beta_j y)^{2s} \tag{1-3}$$

for some $\lambda \in \mathbb{C}$.

The somewhat unsatisfactory nature of the asymmetric summand in (1-3) has been the inspiration for other canonical forms for binary forms of even degree.

Another familiar canonical form is the generalization of (1-1) into the upper-triangular expression for quadratic forms, found by repeated completion of the square:

Theorem 1.2. *A general quadratic form $p \in H_2(\mathbb{C}^n)$ can be written as*

$$p(x_1, \dots, x_n) = \sum_{k=1}^{n} (t_{k,k}x_k + t_{k,k+1}x_{k+1} + \cdots + t_{k,n}x_n)^2, \quad t_{k,l} \in \mathbb{C}. \tag{1-4}$$

The expression in (1-4) is unique, up to the signs of the linear forms.

There are two ways to verify that a candidate expression $F(t; x)$ is, in fact, a canonical form. One is the classical nonconstructive method based on the existence of a point at which the Jacobian matrix has full rank. (See Corollary 2.3, and see Theorem 3.2 for the apolar version.) Lasker [1904] attributes the underlying idea to Kronecker and Lüroth — see [Wakeford 1920, p. 208].

Ideally, however, a canonical form can be derived constructively, and the number of different representations can thereby be determined. The convention in this paper will be that two representations are the same if they are equal, up to a permutation of like summands and with the identification of f^k and $(\zeta f)^k$ when $\zeta^k = 1$. The representation in (1-2) is unique in this sense, even though there are $s! \cdot (2s-1)^s$ different $2s$-tuples $(\alpha_1, \beta_1, \dots, \alpha_s, \beta_s)$ for which (1-2) is valid.

In addition to Theorem 1.1, another motivational example for this paper is a remarkable canonical form for cubic forms found in [Reichstein 1987], which can be thought of as a "completion of the cube".

Theorem 1.3 (Reichstein). *A general cubic $p \in H_3(\mathbb{C}^n)$ can be written uniquely as*

$$p(x_1, \dots, x_n) = \sum_{k=1}^{n} l_k^3(x_1, \dots, x_n) + q(x_3, \dots, x_n), \tag{1-5}$$

where $l_k \in H_1(\mathbb{C}^n)$ and $q \in H_3(\mathbb{C}^{n-2})$.

This is a canonical form, provided q is viewed as a t-linear combination of the monomials in $(x_3, \dots, x_n)$; since $N(n, 3) = n^2 + N(n-2, 3)$, the constant count is right. Iteration (see (6-1)) gives p as a sum of roughly $n^2/4$ cubes. The minimum from constant-counting, which is justified by the Alexander–Hirschowitz theorem [1995], is roughly $n^2/6$. We give Reichstein's constructive proof of Theorem 1.3 in Section 6.

Here are some representative examples of the new canonical forms in this paper.

Theorem 1.4. *A general cubic form $p \in H_3(\mathbb{C}^n)$ has a unique representation*

$$p(x_1, \dots, x_n) = \sum_{1 \le i \le j \le n} (t_{\{i,j\},i} x_i + \cdots + t_{\{i,j\},j} x_j)^3, \tag{1-6}$$

where $t_{\{i,j\},k} \in \mathbb{C}$.

Theorem 1.5. *A general binary sextic $p \in H_6(\mathbb{C}^2)$ can be written as $p(x, y) = f^2(x, y) + g^3(x, y)$, where $f \in H_3(\mathbb{C}^2)$ is a cubic form and $g \in H_2(\mathbb{C}^2)$ is a quadratic form.*

Theorem 1.4 has a constructive proof. Theorem 1.5 is, in fact, a very special case of much deeper recent results of Várilly-Alvarado [2008, especially Theorem 1.2 and Remark 4.5; 2011, Section 1.2]). We include it because our proof, in the next section, is very short.

Theorems 1.1 and 1.5 are both special cases of a more general class of canonical forms for $H_d(\mathbb{C}^2)$, which is a corollary of [Ehrenborg and Rota 1993, Theorem 4.4] (see Theorem 3.4), but not worked out explicitly there.

Theorem 1.6. *Suppose $d \ge 1$, $\{l_j : 1 \le j \le m\}$ is a fixed set of pairwise non-proportional linear forms, and suppose $e_k \mid d$, $d > e_1 \ge \cdots \ge e_r$, $1 \le k \le r$, and*

$$m + \sum_{k=1}^{r} (e_k + 1) = d + 1. \tag{1-7}$$

Then a general binary d-ic form $p \in H_d(\mathbb{C}^2)$ can be written as

$$p(x, y) = \sum_{j=1}^{m} t_j l_j^d(x, y) + \sum_{k=1}^{r} f_k^{d/e_k}(x, y), \tag{1-8}$$

where $t_j \in \mathbb{C}$ and $\deg f_k = e_k$.

The condition $e_k < d$ excludes the vacuous case $m = 0$, $r = 1$, $e_1 = d$. If each $e_k = 1$ and $r = \lfloor (d+1)/2 \rfloor$, then $m = d+1-2\lfloor (d+1)/2 \rfloor \in \{0, 1\}$ and Theorem 1.6 becomes Theorem 1.1; Theorem 1.5 is Theorem 1.6 in the special case $d = 6$, $m = 0$, $r = 2$, $e_1 = 3$, $e_2 = 2$. As an example of a canonical form that is unlikely to find a constructive proof: for a general $p \in H_{84}(\mathbb{C}^2)$, there exist $f \in H_{42}(\mathbb{C}^2)$, $g \in H_{28}(\mathbb{C}^2)$ and $h \in H_{12}(\mathbb{C}^2)$ so that $p = f^2 + g^3 + h^7$.

By taking $d = 2s$, $e_1 = 2$, $e_2 = \cdots = e_{s-1} = 1$ and $m = 0$, in Theorem 1.6, we obtain an alternative to the dangling term λx^{2s} in (1-3).

Corollary 1.7. *A general binary form p of even degree $2s$ can be written as*

$$p(x, y) = (\alpha_0 x^2 + \beta_0 xy + \gamma_0 y^2)^s + \sum_{j=1}^{s-1} (\alpha_j x + \beta_j y)^{2s}. \tag{1-9}$$

A different generalization of Theorem 1.1 focuses on the number of summands.

Theorem 1.8. *A general binary form of degree uv can be written as a sum of $\lceil (uv+1)/(u+1) \rceil$ v-th powers of binary forms of degree u.*

Cayley proved that, after an invertible linear change of variables $(x, y) \mapsto (X, Y)$, a general binary quartic can be written as $X^4 + 6\lambda X^2 Y^2 + Y^4$. There are two natural ways to generalize this to higher even degree, and, almost 100 years ago, Wakeford [1913; 1920] did both.

Theorem 1.9 (Wakeford's theorem). *After an invertible linear change of variables, a general $p \in H_d(\mathbb{C}^n)$ can be written so that the coefficient of each x_i^d is 1 and the coefficient of each $x_i^{d-1} x_j$ is 0.*

There are $N(n, d) - n^2$ unmentioned monomials above, and, when combined with the n^2 coefficients in the change of variables, the constant count is correct for a canonical form. Wakeford was also interested in knowing *which* sets of $n(n-1)$ monomials can be eliminated by a change of variables, and we are able to settle this for binary forms in Theorem 2.4. (Theorem 1.9 was independently discovered in [Guazzone 1975], as an attempt to generalize the canonical form $X^3 + Y^3 + Z^3 + 6\lambda XYZ$ for $H_3(\mathbb{C}^3)$. Babbage [1976] subsequently observed that this can be proved by the Lasker–Wakeford theorem, without noting that Wakeford [1920] had already done so.)

The second generalization of $X^4 + 6\lambda X^2 Y^2 + Y^4$ will not be pursued here; see [Ehrenborg and Rota 1993, Corollary 4.11]. A canonical form for binary forms of even degree $2s$ is given by

$$\sum_{k=1}^{s} l_k^{2s}(x, y) + \lambda \prod_{k=1}^{s} l_k^2(x, y), \quad l_k(x, y) = \alpha_k x + \beta_k y. \tag{1-10}$$

This construction is due to Sylvester [1851b] for $2s = 4, 8$. His methods failed for $2s = 6$, but Wakeford [1913] was able to prove it. The full version of (1-10) is proved in [Wakeford 1920, p. 408], where he notes that "the number of ways this reduction can be performed is interesting", citing "3, 8, 5" for $2s = 4, 6, 8$.

The nontrivial study of canonical forms was initiated by Clebsch's discovery [1861] (see, e.g., [Geramita 1996, pp. 50–51; Reznick 1992a, pp. 59–60]) that, despite the fact that $N(3,4) = 5 \times N(3,1)$, a general ternary quartic cannot be written as a sum of five fourth powers of linear forms. This was early evidence that constant-counting can fail. But $N(3,4)$ is also equal to $1 \times N(3,2) + 3 \times N(3,1)$, and ternary quartics *do* satisfy an alternative canonical form as a mixed sum of powers.

Theorem 1.10. *A general ternary quartic $p \in H_4(\mathbb{C}^3)$ can be written as*

$$p(x_1,x_2,x_3) = q^2(x_1,x_2,x_3) + \sum_{k=1}^{3} l_k^4(x_1,x_2,x_3), \tag{1-11}$$

where $q \in H_2(\mathbb{C}^3)$ and $l_k \in H_1(\mathbb{C}^3)$.

As an alternative generalization of canonical forms, one might also consider polynomial maps $F : S \mapsto H_d(\mathbb{C}^n)$, where S is an N-dimensional subspace of some $\mathbb{C}^M$. In the simplest case, for binary quadratic forms, observe that the coefficient of x^2 in

$$(t_1x + t_2y)^2 + (it_1x + t_3y)^2 \tag{1-12}$$

is 0, so (1-12) is not canonical. This is essentially the only kind of exception.

Theorem 1.11. *Suppose $(c_1,c_2,c_3,c_4) \in \mathbb{C}^4$, and it is not true that $c_3 = \epsilon c_1$ and $c_4 = \epsilon c_2$ for $\epsilon \in \{\pm i\}$. Then, for general $p \in H_2(\mathbb{C}^2)$, there exists $(t_1,t_2,t_3,t_4) \in \mathbb{C}^4$ satisfying*

$$\sum_{j=1}^{4} c_j t_j = 0$$

and such that

$$p(x,y) = (t_1x + t_2y)^2 + (t_3x + t_4y)^2. \tag{1-13}$$

In the exceptional case, there exists (x_0,y_0) so that, for all feasible choices of t_j, $p(x_0,y_0) = 0$.

Another alternative version of (1-3) is the following conjecture, which can be verified up to degree 8.

Conjecture 1.12. *A general binary form p of even degree $2s$ can be written as*

$$p(x,y) = \sum_{j=1}^{s+1} (\alpha_j x + \beta_j y)^{2s}, \quad \text{where } \sum_{j=1}^{s+1} (\alpha_j + \beta_j) = 0. \tag{1-14}$$

Outline. Here is an outline of the paper. In Section 2, we introduce notation and definitions. The definition of canonical form is the classical one and roughly parallels that in [Ehrenborg and Rota 1993], an important updating of this subject about 20 years ago. Our point of view is considerably more elementary in many respects than [Ehrenborg and Rota 1993], but uses the traditional criterion: a polynomial map $F : \mathbb{C}^N \mapsto H_d(\mathbb{C}^n)$ is a *canonical form* if a general $p \in H_d(\mathbb{C}^n)$ is in the range; this occurs if and only if there is at least one point $u \in \mathbb{C}^N$ so that $\{\partial F/\partial t_j(u)\}$ spans $H_d(\mathbb{C}^n)$. (See Corollary 2.3.) This leads to immediate nonconstructive proofs of Theorems 1.2, 1.5, 1.9 and 1.10, and a somewhat more complicated proof of Theorem 2.4, which answers Wakeford's question about missing monomials for binary forms.

In Section 3, we discuss classical apolarity and its implications for canonical forms. (Apolarity methods become more complicated when a component of a canonical form comes from a restricted set of monomials.) A generalization of the classical fundamental theorem of apolarity from [Reznick 1996] allows us to identify a class of bases for $H_d(\mathbb{C}^n)$ which give a nonconstructive proof of Theorem 1.6, and hence Theorem 1.1. A similar argument yields the proof of Theorem 1.8. We also present Sylvester's algorithm, Theorem 3.8, allowing for a constructive proof of Theorem 1.1. We conclude with a brief summary of connections with the theorems of Alexander and Hirschowitz and recent work on the rank of forms.

In Section 4 we discuss some special cases of Theorem 1.6. Sylvester's algorithm is used in constructive proof of Theorem 1.6 when $e_k \equiv 1$, in which case the representation is unique. We give some other constructive proofs for $d \le 4$, and present numerical evidence regarding the number of representations in Corollary 1.7 and a few other cases. Using elementary number theory, we show that, for each r, there are only finitely many canonical forms (1-8) with $m = 0$, and, up to degree N, there are $N + \mathbb{O}(N^{1/2})$ such canonical forms in which the e_k are equal.

Section 5 discusses some familiar results on sums of two squares of binary forms and canonical representations of quadratic forms as a sum of squares of linear forms. This includes a constructive proof of Theorem 1.2, which provides the groundwork for the proof of Theorem 1.4. We also give a short proof of a canonical form which illustrates the classical result that a general ternary quartic is the sum of three squares of quadratic forms.

In Section 6, we turn to forms in more than two variables and low degree, give constructive proofs of Theorems 1.3 and 1.4, as well as the noncanonical Theorem 6.2, which shows that *every* cubic in $H_3(\mathbb{C}^n)$ is a sum of at most $n(n+1)/2$ cubes of linear forms. Theorem 1.3 can be "lifted" to an ungainly canonical form for quartics as a sum of fourth powers (see Corollary 6.3), but not further to quintics. Number theoretic considerations rule out a Reichstein-type canonical form for quartics in 12 variables; see Proposition 6.4 for other instances of this phenomenon.

In Section 7, we offer a preliminary discussion of canonical forms in which the domain of a polynomial map $F : \mathbb{C}^M \mapsto H_d(\mathbb{C}^n)$ is restricted to an N-dimensional subspace of $\mathbb{C}^M$, of which Theorem 1.11 and Conjecture 1.12 are examples.

2. Basic definitions, and proofs of Theorems 1.2, 1.5, 1.9 and 1.10

Let $\mathcal{I}(n, d)$ denote the index set of monomials in $H_d(\mathbb{C}^n)$:

$$\mathcal{I}(n, d) = \Big\{(i_1, \dots, i_n) : 0 \le i_k \in \mathbb{Z}, \sum_k i_k = d\Big\}. \tag{2-1}$$

Let $x^i = x_1^{i_1} \cdots x_n^{i_n}$ and $c(i) = d!/\big(\prod i_k!\big)$ denote the multinomial coefficient. If $p \in H_d(\mathbb{C}^n)$, then we write

$$p(x_1, \dots, x_n) = \sum_{i \in \mathcal{I}(n,d)} c(i)a(p; i)x^i, \quad a(p; i) \in \mathbb{C}. \tag{2-2}$$

We say that two forms are *distinct* if they are nonproportional, and a set of forms is *honest* if the forms are pairwise distinct. For later reference, recall Biermann's theorem; see [Reznick 1992a, p. 31].

Theorem 2.1 (Biermann's theorem). *If $p \in H_d(\mathbb{C}^n)$ and $p \neq 0$, then there exists $i \in \mathcal{I}(n, d)$ so that $p(i) \neq 0$.*

The easy verification of whether a formula is a canonical form for $H_d(\mathbb{C}^n)$ relies on a crucial alternative. A self-contained accessible proof is in [Ehrenborg and Rota 1993, Theorem 2.4], for which Ehrenborg and Rota thank M. Artin and A. Mattuck. For further discussion of the underlying algebraic geometry, see Section 9.5 in [Cox et al. 2007].

Theorem 2.2. *Suppose $M \ge N$ and $F : \mathbb{C}^M \to \mathbb{C}^N$ is a polynomial map*; *that is*,

$$F(t_1, \dots, t_M) = \big(f_1(t_1, \dots, t_M), \dots, f_N(t_1, \dots, t_M)\big),$$

where each $f_j \in \mathbb{C}[t_1, \dots, t_M]$. Then either (i) *or* (ii) *holds*:

(i) *The N polynomials $\{f_j : 1 \le j \le N\}$ are algebraically dependent and $F(\mathbb{C}^M)$ lies in some nontrivial variety $\{P = 0\}$ in $\mathbb{C}^N$.*

(ii) *The N polynomials $\{f_j : 1 \le j \le N\}$ are algebraically independent and $F(\mathbb{C}^M)$ is dense in $\mathbb{C}^N$.*

The second case occurs if and only there is a point $u \in \mathbb{C}^M$ at which the Jacobian matrix $[\partial f_i/\partial t_j(u)]$ has full rank.

When $M = N = N(n, d)$, we may interpret such an F as a map from $\mathbb{C}^N$ to $H_d(\mathbb{C}^n)$ by indexing $\mathcal{I}(n, d)$ as $\{i(k) : 1 \le k \le N\}$ and making the interpretation in an abuse of notation that

$$F(t;x)=\sum_{k=1}^{N} c(i(k))\, f_k(t_1\dots,t_N)x^{i(k)}. \tag{2-3}$$

Definition. A *canonical form for* $H_d(\mathbb{C}^n)$ is any polynomial map $F:\mathbb{C}^{N(n,d)}\mapsto H_d(\mathbb{C}^n)$ in which F satisfies Theorem 2.2(ii).

That is, F is a canonical form if and only if $N=N(n,d)$ and, for a general $p\in H_d(\mathbb{C}^n)$, there exists $t\in\mathbb{C}^N$ so that $p(x)=F(t;x)$. The significance of this choice of N is that it is the smallest possible value. In the rare cases where F is surjective, we say that the canonical form is *universal*.

By translating the definitions and using (2-1) and (2-3), we obtain an immediate corollary of Theorem 2.2:

Corollary 2.3. *The polynomial map* $F:\mathbb{C}^N\mapsto H_d(\mathbb{C}^n)$ *is a canonical form if and only if there exists* $u\in\mathbb{C}^n$ *so that* $\{\partial F/\partial t_j(u)\}$ *spans* $H_d(\mathbb{C}^n)$.

We shall let $J:=J(F;u)$ denote the span of the forms $\{\partial F/\partial t_j(u)\}$. In any particular case, the determination of whether $J=H_d(\mathbb{C}^n)$ amounts to the computation of the determinant of an $N(n,d)\times N(n,d)$ matrix. As much as possible in this paper, we give proofs which can be checked by hand, by making a judicious choice of u and ordering of the monomials in $H_d(\mathbb{C}^n)$, showing sequentially that they all lie in J.

Classically, the use of the term "canonical form" has been limited to cases in which $F(t;x)$ has a natural interpretation as a combination of forms in $H_d(\mathbb{C}^n)$, such as a sum of powers of linear forms, or as a result of a linear change of variables. It seems odd that canonical forms are perceived as rare, since a "general" polynomial map from $\mathbb{C}^N\mapsto H_d(\mathbb{C}^n)$ is a canonical form. (This is an observation which goes back at least to [Richmond 1902].) For example, if $\{f_j(x)\}$ is a basis for $H_d(\mathbb{C}^n)$, then

$$F(t;x)=\sum_{j=1}^{N} t_j f_j(x) \tag{2-4}$$

should be (but usually isn't) considered a canonical form. In particular, (2-2) with $f_j(x)=c(i_j)x^{i_j}$ is itself a canonical form.

The following computation will occur repeatedly. If $es=d$, then

$$g=\sum_{i_j\in\mathcal{I}(n,e)} t_j x^{i_j} \implies \frac{\partial g^s}{\partial t_j}=s x^{i_j} g^{s-1}. \tag{2-5}$$

If g is specialized to be a monomial, then all these partials will also be monomials.

Nonconstructive proof of Theorem 1.2. Given (1-4), let

$$l_k(x) = \sum_{m=k}^{n} t_{k,m} x_m, \quad F(x) = \sum_{k=1}^{n} l_k^2(x).$$

Then $\partial F/\partial t_{k,m} = 2x_m l_k$. Set $t_{k,m} = \delta_{k,m}$, so that $l_k = x_k$ and $\partial F/\partial t_{k,m} = 2x_k x_m$. Since $1 \le k \le m \le n$, all monomials from $H_2(\mathbb{C}^n)$ appear in J. $\square$

Nonconstructive proof of Theorem 1.5. Suppose

$$p(x, y) = f^2(x, y) + g^3(x, y), \tag{2-6}$$

with

$$f(x, y) = t_1 x^3 + t_2 x^2 y + t_3 x y^2 + t_4 y^3, \quad g(x, y) = t_5 x^2 + t_6 x y + t_7 y^2.$$

Then, by (2-5), the partials with respect to the t_j are

$$2x^3 f,\ 2x^2 y f,\ 2xy^2 f,\ 2y^3 f; \quad 3x^2 g^2,\ 3xyg^2,\ 3y^2 g^2.$$

Upon specializing at $f = x^3$, $g = y^2$, these become

$$2x^6,\ 2x^5 y,\ 2x^4 y^2,\ 2x^3 y^3; \quad 3x^2 y^4,\ 3xy^5\ , 3y^6.$$

It is then evident that $J = H_6(\mathbb{C}^2)$. $\square$

Nonconstructive proof of Theorem 1.9. Let $\mathcal{L} \subset \mathcal{I}(n, d)$ consist of all n-tuples except the permutations of $(d, 0, \dots, 0)$ and $(d-1, 1, \dots, 0)$ and let $X_i = \sum_{j=1}^{n} \alpha_{ij} x_j$. The assertion is that, with the $\bigl(N(n, d) - n - \binom{n}{2}\bigr) + n^2 = N(n, d)$ parameters t_l and α_{ij},

$$\sum_{i=1}^{n} X_i^d + \sum_{l \in \mathcal{L}} t_l X_1^{l_1} \cdots X_n^{l_n} \tag{2-7}$$

is a canonical form. Evaluate the partials at the point where $X_i = x_i$ and $t_l = 0$: they are $d x_j x_i^{d-1}$ (for α_{ij}) and x^l (for t_l). Taking $1 \le i, j \le n$ and $l \in \mathcal{L}$, we see that J contains all monomials in $H_d(\mathbb{C}^n)$. $\square$

As a special case (used later in Theorem 4.6), we obtain the familiar result that, after appropriate linear changes of variable, a general binary quartic may be written as $x^4 + 6\lambda x^2 y^2 + y^4$. It is classically known (see [Elliott 1913, Section 211]) that the choice of λ is not unique: in fact, after appropriate linear changes of variable, $x^4 + 6\lambda x^2 y^2 + y^4$ can be written as $x^4 + 6\mu x^2 y^2 + y^4$ for $\mu \in \{\pm\lambda, \pm(1-\lambda)/(1+3\lambda), \pm(1+\lambda)/(1-3\lambda)\}$.

Wakeford asserts that Theorem 1.9 is also true with $x_i^{d-1} x_j$ replaced by $x_i^{d-r} x_j^r$ (evidently when $r \ne d/2$), but his proof seems sketchy. He also gives necessary conditions for sets of $n(n-1)$ monomials which may be omitted, and these are hard to follow as well. Below, we answer his question in the binary case: in the only two excluded cases below, (2-8) has a square factor, and so cannot be canonical.

Theorem 2.4. *Let $\mathcal{B} = (m_1, m_2, n_1, n_2)$ be four distinct integers in $\{0, \dots, d\}$ so that $\{m_1, m_2\} \neq \{0, 1\}, \{d-1, d\}$. Then, after an invertible linear change of variable, a general binary form p of degree d can be written as*

$$p(x, y) = x^{d-n_1} y^{n_1} + x^{d-n_2} y^{n_2} + \sum_{k \notin \mathcal{B}} t_k x^{d-k} y^k \quad \text{for some } \{t_k\} \subset \mathbb{C}. \tag{2-8}$$

Proof. Writing $(x, y) \mapsto (\alpha_1 x + \alpha_2 y, \alpha_3 x + \alpha_4 y) := (X, Y)$, we have

$$F = X^{d-n_1} Y^{n_1} + X^{d-n_2} Y^{n_2} + \sum_{k \notin \mathcal{B}} t_k X^{d-k} Y^k. \tag{2-9}$$

Evaluate the partials of (2-9) at $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (1, 0, 0, 1)$ (so $X = x$, $Y = y$) and $t_k = 1$ (note the difference with the previous proof, in which $t_k = 0$). The $d - 3$ partials with respect to the t_k are simply $x^{d-k} y^k$, $k \notin \mathcal{B}$, so these are in J. Further,

$$\frac{\partial F}{\partial \alpha_1} = \sum_{j \neq m_1, m_2} (d - j) x^{d-j} y^j, \qquad \frac{\partial F}{\partial \alpha_4} = \sum_{j \neq m_1, m_2} j x^{d-j} y^j. \tag{2-10}$$

Since most monomials used in (2-10) are already in J, it follows that J also contains

$$(d - n_1) x^{d-n_1} y^{n_1} + (d - n_2) x^{d-n_2} y^{n_2}, \quad n_1 x^{d-n_1} y^{n_1} + n_2 x^{d-n_2} y^{n_2}, \tag{2-11}$$

and since $(d - n_1) n_2 \neq (d - n_2) n_1$, (2-11) implies that $x^{d-n_j} y^{n_j} \in J$ for $j = 1, 2$. To this point, we have shown that J contains all monomials from $H_d(\mathbb{C}^2)$ except for $x^{d-m_j} y^{m_j}$, where $m_1 < m_2$. The two remaining partial derivatives are

$$\frac{\partial F}{\partial \alpha_2} = \sum_{j \neq m_1, m_2} (d - j) x^{d-j-1} y^{j+1}, \qquad \frac{\partial F}{\partial \alpha_3} = \sum_{j \neq m_1, m_2} j x^{d-j+1} y^{j-1}, \tag{2-12}$$

and so J contains as well the forms in (2-12) of the shape $c_1 x^{d-m_1} y^{m_1} + c_2 x^{d-m_2} y^{m_2}$. We need to distinguish a number of cases. If $m_1 = 0$, $m_2 = d$, then these forms are y^d, x^d. If $m_1 = 0$ and $2 \le m_2 \le d - 1$, then these forms are $(d - m_2) x^{d-m_2} y^{m_2}$ and $x^d + (m_2 + 1) x^{d-m_2} y^{m_2}$, and similarly when $1 \le m_1 \le d - 2$ and $m_2 = d$. (Recall that we have excluded the cases $(m_1, m_2) = (0, 1)$ and $(d - 1, d)$). In the remaining cases, $1 \le m_1 < m_2 \le d - 1$. If $m_2 = m_1 + 1$, then these forms are $(d - (m_1 - 1)) x^{d-m_1} y^{m_1}$ and $(m_2 + 1) x^{d-m_2} y^{m_2}$. Finally, if $m_2 > m_1 + 1$, then all four terms appear, and the forms are

$$\begin{gathered}(d - m_1 + 1) x^{d-m_1} y^{m_1} + (d - m_2 + 1) x^{d-m_2} y^{m_2}, \\ (m_1 + 1) x^{d-m_1} y^{m_1} + (m_2 + 1) x^{d-m_2} y^{m_2}.\end{gathered} \tag{2-13}$$

In each of the cases, linear combinations of the forms produce the missing monomials, so $J = H_d(\mathbb{C}^2)$. ☐

Remark. By writing $p(x, y) = \prod_k (x + \alpha_k y)$, it follows from Theorem 1.9 that, for a general set of d complex numbers α_k, there exists a Möbius transformation T so that

$$\sum_{k=1}^{d} T(\alpha_k) = 0, \quad \sum_{k=1}^{d} T\left(\frac{1}{\alpha_k}\right) = 0, \quad \prod_{k=1}^{d} T(\alpha_k) = 1. \tag{2-14}$$

Nonconstructive proof of Theorem 1.10. Write (1-11) as $F(x; t)$, where

$$q(x_1, x_2, x_3) = t_1 x_1^2 + t_2 x_2^2 + t_3 x_3^2 + t_4 x_1 x_2 + t_5 x_1 x_3 + t_6 x_2 x_3,$$
$$l_k(x_1, x_2, x_3) = t_{k1} x_1 + t_{k2} x_2 + t_{k3} x_3.$$

Evaluate the partials at $q = x_1x_2 + x_1x_3 + x_2x_3$ and $(l_1, l_2, l_3) = (x_1, x_2, x_3)$. Then $\partial F/\partial t_{kl} = 4x_l x_k^3$, so $x_i^4, x_i^3 x_j \in J$; since $\partial F/\partial t_1 = 2x_1^2 q = 2x_1^2(x_1x_2+x_1x_3+x_2x_3)$, it follows that $x_1^2 x_2 x_3 \in J$. Similarly, by considering $\partial F/\partial t_2$ and $\partial F/\partial t_3$, it follows that $x_1 x_2^2 x_3$, $x_1 x_2 x_3^2$ are in J. Finally, $\partial F/\partial t_4 = 2x_1x_2 q = 2x_1x_2(x_1x_2+x_1x_3+x_2x_3)$, and so now $x_1^2 x_2^2 \in J$. Similarly, by considering $\partial F/\partial t_5$ and $\partial F/\partial t_6$, it follows that $x_1^2 x_3^2$, $x_2^2 x_3^2$ are also in J, and this accounts for all monomials in $H_4(\mathbb{C}^3)$. □

Other applications of Corollary 2.3 to canonical forms can be found in [Ehrenborg and Rota 1993], including interpretations of the older results in [Richmond 1902; Turnbull 1960, pp. 265–269].

3. Apolarity and proofs of Theorems 1.1, 1.6 and 1.8

Using the notation of (2-1) and (2-2), for $p, q \in H_d(\mathbb{C}^n)$, define the following bilinear form:

$$[p, q] = \sum_{i \in \mathcal{I}(n,d)} c(i) a(p; i) a(q; i). \tag{3-1}$$

Recall two basic notations. For $\alpha \in \mathbb{C}^n$, define $(\alpha \cdot)^d \in H_d(\mathbb{C}^n)$ by

$$(\alpha \cdot)^d(x) = (\alpha \cdot x)^d = \left(\sum_{j=1}^{n} \alpha_j x_j\right)^d = \sum_{i \in \mathcal{I}(n,d)} c(i) \alpha^i x^i. \tag{3-2}$$

Define the differential operator $f(D)$ for $f \in H_e(\mathbb{C}^n)$ in the usual way by

$$f(D) = \sum_{i \in \mathcal{I}(n,e)} c(i) a(f; i) \left(\frac{\partial}{\partial x_1}\right)^{i_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{i_n}. \tag{3-3}$$

It follows immediately that, for $\alpha \in \mathbb{C}^n$,

$$[p, (\alpha \cdot)^d] = \sum_{i \in \mathcal{I}(n,d)} c(i) a(p; i) \alpha^i = p(\alpha). \tag{3-4}$$

If $i \neq j \in \mathcal{I}(n,d)$, then $i_k > j_k$ for some k, so $D^i x^j = 0$; otherwise $D^i x^i = \prod_k (i_k)! = d!/c(i)$. Suppose $p, q \in H_d(\mathbb{C}^n)$. Bilinearity and (3-3) imply the classical result that

$$
\begin{aligned}
\text{(3-5)} \qquad p(D)q &= \sum_{i \in \mathcal{I}(n,d)} c(i)a(p;i)D^i \Bigl(\sum_{j \in \mathcal{I}(n,d)} c(j)a(q;j)x^j \Bigr) \\
&= \sum_{i \in \mathcal{I}(n,d)} \sum_{j \in \mathcal{I}(n,d)} c(i)c(j)a(p;i)a(q;j)D^i x^j \\
&= \sum_{i \in \mathcal{I}(n,d)} c(i)c(i)a(p;i)a(q;i)D^i x^i \\
&= \sum_{i \in \mathcal{I}(n,d)} c(i)^2 a(p;i)a(q;i)\frac{d!}{c(i)} = d!\,[p,q] = d!\,[q,p] = q(D)p.
\end{aligned}
$$

Definition. If $p \in H_d(\mathbb{C}^n)$ and $q \in H_e(\mathbb{C}^n)$, then p and q are *apolar* if $p(D)q = q(D)p = 0$.

Note that, if $d = e$, then p and q are apolar if and only if $[p,q] = 0$, and, if $d > e$, say, then the equation $p(D)q = 0$ is automatic, so only $q(D)p = 0$ need be checked. By (3-4), p is apolar to $(\alpha \cdot)^d$ if and only if $p(\alpha) = 0$.

The following lemma is both essential and trivial.

Lemma 3.1. *Suppose $X = \operatorname{span}(\{h_j\}) \subseteq H_d(\mathbb{C}^n)$. Then $X = H_d(\mathbb{C}^n)$ if and only if there is no $0 \neq p \in H_d(\mathbb{C}^n)$ which is apolar to each of the h_j.*

From this point of view, Theorem 3.2 is a direct consequence of Corollary 2.3:

Theorem 3.2 (Lasker–Wakeford). *If $F : \mathbb{C}^N \to H_d(\mathbb{C}^n)$, then F is a canonical form if and only if there is a point u so that there is no nonzero form $q \in H_d(\mathbb{C}^n)$ which is apolar to all N forms $\{\partial F/\partial t_k(u)\}$.*

The attribution "Lasker–Wakeford" (for [Lasker 1904; Wakeford 1920]) is taken from [Turnbull 1960]: H. W. Turnbull (1885–1961) was one of the last practicing invariant theorists who had been trained in the pre-Hilbert approach; see [Fisher 1967, pp. 231–232]. (His text [Turnbull 1960] is a Rosetta Stone for understanding the 19th century approach to algebra in more modern terminology.) Turnbull referred to Theorem 3.2 as "paradoxical and very curious". E. Lasker (1868–1941) received his Ph.D. under M. Noether at Göttingen in 1902. He is probably better known for being the world chess champion for 27 years (1894–1921), spanning the life of E. K. Wakeford (1894–1916). J. H. Grace, Wakeford's professor at Oxford, edited the second half of his thesis into [Wakeford 1920] and also wrote a memorial article [Grace 1918] for him in 1918:

> "He [EKW] was slightly wounded early in 1916, and soon after coming home was busy again with Canonical Forms... [H]e discovered a paper

of Hilbert's which contained the very theorem he had long been in want of — first vaguely, and later quite definitely. This was in March; April found him, full of the most joyous and reverential admiration for the great German master, working away in fearful haste to finish the dissertation... He returned to the front in June and was killed in July... He only needed a chance, and he never got it."

The following properties are easily established; see, e.g., [Reznick 1992a; 1996] for proofs.

Theorem 3.3. (i) *If $e \le d$ and $f \in H_e(\mathbb{C}^n)$, $g \in H_{d-e}(\mathbb{C}^n)$ and $p \in H_d(\mathbb{C}^n)$, then*

$$d!\,[fg, p] = (fg)(D)p = f(D)g(D)p = e!\,[f, g(D)p]. \tag{3-6}$$

Thus, p is apolar to every multiple of g in $H_d(\mathbb{C}^n)$ if and only if p and g are apolar.

(ii) *If $p \in H_d(\mathbb{C}^n)$, then $(1/d)\partial p/\partial x_j(\alpha) = [p, x_j(\alpha\cdot)^{d-1}]$. Thus, p is apolar to $(\alpha\cdot)^{d-1}$ if and only if p is singular at α. More generally, p is apolar to $(\alpha\cdot)^{d-e}$ if and only if p vanishes to e-th order at α.*

(iii) *If $e \le d$ and $g \in H_{d-e}(\mathbb{C}^n)$, then $g(D)(\alpha\cdot)^d = (d!/e!)g(\alpha)(\alpha\cdot)^e$.*

Suppose $F(t; x)$ contains h^s as a summand, where $h(x) = \sum_{l\in\mathcal{I}(n,e)} t_l x^l$, and suppose that no t_l occurs elsewhere in $F(t; x)$. If p is apolar to each partial of F, then it will be apolar to $\partial F/\partial t_l = s x^l h^{s-1}$ by (2-5). Since this is true for every $l \in \mathcal{I}(n, e)$, it follows from (i) that p is apolar to h^{s-1}. It is critical to note that this observation requires that each of the monomials of degree e appear in h, and does not apply if h is defined as a sum from a restricted set of monomials.

We are now able to give a short proof of the "second main theorem on apolarity" from [Ehrenborg and Rota 1993], which was not concerned with preserving the constant count.

Theorem 3.4. *Suppose $j_l = (j_{l,1}, \dots, j_{l,m})$, $1 \le l \le r$, are m-tuples of nonnegative integers, and suppose positive integers d_k, $1 \le k \le m$, and d are chosen so that*

$$u_l := d - \sum_{k=1}^{m} j_{l,k} d_k \ge 0 \tag{3-7}$$

for each l. Fix forms $q_l \in H_{u_l}(\mathbb{C}^n)$ and, for $f_k \in H_{d_k}(\mathbb{C}^n)$, define

$$F(f_1, \dots, f_m) = \sum_{l=1}^{r} q_l(x) f_1^{j_{l,1}} \cdots f_m^{j_{l,m}}. \tag{3-8}$$

Let $F_j := \partial F/\partial f_j$. Then a general $p \in H_d(\mathbb{C}^n)$ can be written as (3-8) if and only if there exists a specific $\bar f = (\bar f_k)$ so that no nonzero $p \in H_d(\mathbb{C}^n)$ is apolar to *each* $F_j(\bar{f})$, $1 \le j \le m$. *If, in addition,*

$$\sum_{k=1}^{m} N(n, d_k) = N(n, d), \tag{3-9}$$

then (3-8) is a canonical form.

Proof. Let

$$f_j(x) = \sum_{i_v \in \mathcal{I}(n,d_j)} t_{j,v} x^{i_v}. \tag{3-10}$$

By Theorem 2.2, (3-7) and Lemma 3.1, (3-8) represents general $p \in H_d(\mathbb{C}^n)$ if and only if there is some $\bar{f}$ so that there is no nonzero form in $p \in H_d(\mathbb{C}^n)$ which is apolar to each $\partial F/\partial t_{j,v}(\bar{f}) = d_k x^{i_v} F_j(\bar{f})$, or, by Theorem 3.3(i), to each $F_j(\bar{f})$. The constant count is checked by (3-9). □

By Theorem 3.3(ii) and Theorem 3.4,

$$F = \sum_{k=1}^{r} (\alpha_k \cdot)^d$$

is a canonical form if and only if there exist r points $\bar{\alpha}_k \in \mathbb{C}^n$ at which no nonzero form $p \in H_d(\mathbb{C}^n)$ is singular. This result is classical, and goes back to [Clebsch 1861]; see also [Ehrenborg and Rota 1993, Theorem 4.2]. A particularly deep result of Alexander and Hirschowitz [1995] states that a general form in $H_d(\mathbb{C}^n)$, $d \ge 3$, may be written as a sum of $\lceil N(n, d)/n \rceil$ d-th powers of linear forms, except when $(n, d) = (5, 3)$, $(3, 4)$, $(4, 4)$, $(5, 4)$, when an extra summand is needed. (For much more on this, see [Geramita 1996, Lecture 7; Iarrobino and Kanev 1999, Corollary 1.62; Landsberg 2012, Chapter 15; Ranestad and Schreyer 2000, Theorem 0.2]; for a brief exposition of the proof, see [Landsberg 2012, Chapter 15].) These references also discuss the exceptional examples, which were all known in the 19th century. The expression of forms as a sum of powers of forms is currently a very active area of interest; see the references above as well as [Carlini et al. 2012; Fröberg et al. 2012; Landsberg and Teitler 2010].

The fundamental theorem of apolarity (see [Reznick 1996] for a history) states that, if f is irreducible and $p \in H_d(\mathbb{C}^n)$, then f and p are apolar if and only if p can be written as a sum of terms of the form $(\alpha_j \cdot)^d$, where $f(\alpha_j) = 0$. This was generalized as follows:

Theorem 3.5 [Reznick 1996, Theorem 4.1]. *Suppose* $q \in H_e(\mathbb{C}^n)$ *factors as* $\prod_{j=1}^r q_j^{m_j}$ *into a product of powers of distinct irreducible factors and suppose* $p \in H_d(\mathbb{C}^n)$. *Then* $q(D)p = 0$ *if and only if there exist* $\alpha_{jk} \subset \{q_j(\alpha) = 0\}$ *and* $\phi_{jk} \in H_{m_j-1}(\mathbb{C}^n)$ *such that*

$$p = \sum_{j=1}^{r} \sum_{k=1}^{n_j} \phi_{jk} (\alpha_{jk} \cdot)^{d-(m_j-1)}.$$

The application of apolarity to binary forms is particularly simple, because zeros correspond to factors. If $e = d + 1$, then $q(D)p = 0$ for every $p \in H_d(\mathbb{C}^n)$, and we obtain the following result, also found in [Ehrenborg and Rota 1993, Theorem 4.5].

Corollary 3.6. *Suppose $\{\alpha_j x + \beta_j y : 1 \le j \le r\}$ is honest and suppose $\sum_{j=1}^r m_j = d + 1$. Then the following set is a basis for $H_d(\mathbb{C}^2)$:*

$$\mathcal{S} = \left\{ x^k y^{m_j - 1 - k} (\beta_j x - \alpha_j y)^{d - m_j + 1} : 0 \le k \le m_j - 1, 1 \le m_j \le r \right\}. \tag{3-11}$$

Proof. If p is apolar to each term in (3-11), then $(\alpha_j x + \beta_j y)^{m_j} \mid p$ by Theorem 3.3(ii). Thus $p = 0$ by degree considerations, and $\mathcal{S}$ has $d + 1$ elements, so it is a basis. □

If each $m_j = 1$, then Corollary 3.6 states that an honest set $\mathcal{S} = \{(\alpha_j x + \beta_j y)^d\}$ of $d + 1$ forms is a basis for $H_d(\mathbb{C}^2)$. This is easily proved directly, since the representation of $\mathcal{S}$ with respect to the basis $\left\{ \binom{d}{j} x^{d-j} y^j \right\}$, $[\alpha_j^{d-k} \beta_j^k]$, has Vandermonde determinant

$$\prod_{1 \le i < j \le n} (\alpha_i \beta_j - \alpha_j \beta_i). \tag{3-12}$$

Each product in (3-12) is nonzero because $\{(\alpha_j x + \beta_j y)^d\}$ is honest. One implication of this independence is found in [Reznick 2013, Corollary 4.3].

Lemma 3.7. *If $p(x, y) \in H_d(\mathbb{C}^2)$ has two honest representations*

$$p(x, y) = \sum_{i=1}^m (\alpha_i x + \beta_i y)^d = \sum_{j=1}^n (\gamma_j x + \delta_j y)^d \tag{3-13}$$

and $m + n \le d + 1$, then the representations are permutations of each other.

Proof. If (3-13) holds, then $\{(\alpha_i x + \beta_i y)^d, (\gamma_j x + \delta_j y)^d\}$ is linearly dependent, which is impossible unless the dependence is trivial. □

It follows immediately from Lemma 3.7 that the representations (1-2) and (1-3), if they exist for p, are unique. When $n \ge 3$, the linear dependence of a set $\{(\alpha_j \cdot)^d\}$ depends on the geometry of the points as well as the number (see the discussion of Serret's theorem in [Reznick 1992a, p. 29].) Even for powers of binary forms of degree $e \ge 2$, there are singular cases. It is not hard to show that a *general* set of $(2k + 1)$ k-th powers of quadratic forms is linearly independent; however, for example, $(x^2 - y^2)^2 + (2xy)^2 = (x^2 + y^2)^2$. For much more on this, see [Reznick $\ge$ 2013].

Nonconstructive proof of Theorem 1.6. For $1 \le k \le r$, write

$$f_k(x, y) = \sum_{l=0}^{e_k} t_{k,l} x^{e_k - l} y^l.$$

By Corollary 2.3 and (2-5), (1-8) is a canonical form in the variables $\{t_j, t_{k,l}\}$ provided there is a point at which the partials

$$\{l_j^d, 1 \le j \le m\} \cup \left\{x^{e_k-l}y^l f_k^{d/e_k-1}, 1 \le l \le e_k, 1 \le k \le r\right\}$$

span $H_d(\mathbb{C}^2)$. Let $f_k = \tilde{l}_k^{e_k}$, where $\{l_1, \dots, l_m, \tilde{l}_1, \dots, \tilde{l}_r\}$ is chosen to be honest. Then, by (1-7), the desired assertion follows immediately from Corollary 3.6. $\square$

Nonconstructive proof of Theorem 1.8. Write $uv+1 = r(u+1)+s$. If $s=0$, then Theorem 1.8 is simply a special case of Theorem 1.6 with $m=0$, $d=uv$ and $e_k \equiv u$. Otherwise, $1 \le s \le u$, so that $r+1 = \lceil (uv+1)/(u+1) \rceil$. Let

$$F(\{\alpha_{ij}\}) = \sum_{i=1}^{r+1} f_i^v(x,y), \qquad f_i(x,y) = \sum_{j=0}^{u} \alpha_{ij}x^{u-j}y^j.$$

This is *not* a canonical form, as there are too many constants. As before,

$$\frac{\partial F}{\partial \alpha_{ij}} = vx^{u-j}y^j f_i^{v-1}.$$

We now specialize to $f_i(x,y) = (ix-y)^u$ and use the apolarity argument to show that $J = H_{uv}(\mathbb{C}^2)$. Suppose $q \in H_{uv}(\mathbb{C}^2)$ is apolar to each partial. Then, by Theorem 3.3, it is apolar to $f_i^{v-1} = (ix-y)^{uv-u}$, and so q vanishes to u-th order at $(i,-1)$ for $1 \le i \le r+1$. It follows that q is a multiple of $\prod_{i=1}^{r+1}(x+iy)^{u+1}$, and so $q=0$ by degree considerations.

It is an exercise to show that F can be converted to an canonical form by requiring, say, that f_{r+1} only contain monomials $x^{u-j}v^j$ for $0 \le j \le s-1$. $\square$

We present now Sylvester's algorithm. For modern discussions of this, along with Gundelfinger's generalization [1887], which is not included here, see [Kung and Rota 1984, Section 5; Kung 1986; 1987; 1990; Reznick 1996; 2013].

Theorem 3.8 (Sylvester's algorithm). *Let*

$$p(x,y) = \sum_{j=0}^{d} \binom{d}{j} a_j x^{d-j} y^j$$

be a given binary form and suppose $\{\alpha_j x + \beta_j y\}$ is honest. Let

$$h(x,y) = \sum_{t=0}^{r} c_t x^{r-t} y^t = \prod_{j=1}^{r} (\beta_j x - \alpha_j y).$$

Then there exist $\lambda_k \in \mathbb{C}$ so that

$$p(x,y) = \sum_{k=1}^{r} \lambda_k (\alpha_k x + \beta_k y)^d$$

if and only if

$$\begin{pmatrix} a_0 & a_1 & \cdots & a_r \\ a_1 & a_2 & \cdots & a_{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d-r} & a_{d-r+1} & \cdots & a_d \end{pmatrix} \cdot \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_r \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3-14}$$

Theorem 3.8 can be put in the context of our previous discussion. Let $A_r(p)$ denote the $(d-r+1)\times(r+1)$ Hankel matrix on the left-hand side of (3-14). If $h(D)=\prod_{j=1}^{r}(\beta_j\partial/\partial x-\alpha_j\partial/\partial y)$, then a direct computation shows that

$$h(D)p=\sum_{m=0}^{d-r}\frac{d!}{(d-r-m)!\,m!}\left(\sum_{i=0}^{d-r}a_{i+m}c_i\right)x^{d-r-m}y^m. \tag{3-15}$$

It follows from (3-15) that the coefficients of $h(D)p$ are thus, up to multiple, the rows of the matrix product, so (3-14) is equivalent to $h(D)p=0$. In this way, Theorem 3.8 follows from Theorem 3.5. Sylvester's algorithm can also be visualized as seeking constant-coefficient linear recurrences satisfied by $\{a_k\}$ and looking for the shortest one whose characteristic equation has distinct roots; this is the proof given in [Reznick 2013]. In this case, Gundelfinger's results handle the case when the roots are not distinct.

Constructive proof of Theorem 1.1. Suppose $d=2s-1$ is odd. The matrix $A_s(p)$ is $s\times(s+1)$ and has a nontrivial null-vector. The corresponding h (which can be given in terms of the coefficients of p) has distinct factors unless its discriminant vanishes. Thus, for general $p\in H_{2s-1}(\mathbb{C}^2)$, Theorem 3.8 gives p as a sum of s $(2s-1)$-st powers of linear forms.

If $d=2s$, the matrix $A_s(p)$ is square, and if p is a sum of s $2s$-th powers, then $\det A_s(p)=0$. Conversely, if $\det A_s(p)=0$ and the corresponding h has distinct factors (which is generally true), then p is a sum of s $2s$-th powers. If M_1 and M_2 are two square matrices and $\operatorname{rank}(M_2)=k$, then $\det(M_1+\lambda M_2)$ is a polynomial in λ of degree k. In particular, if $q=(\alpha x+\beta y)^{2s}$, then $\operatorname{rank}(H_s(q))=1$. Thus, in general, there is a unique value of λ and some matrix M so that $0=\det A_s(p-\lambda(\alpha x+\beta y)^{2s})=\det A_s(p)-\lambda\det M$. (When $\alpha x+\beta y=x$, M is the $(1,1)$-cofactor of $A_s(p)$.) In the special case $\alpha x+\beta y=x$, this proves Theorem 1.1(ii). The same argument shows that, for general $q\in H_{2s}(\mathbb{C}^2)$, there exist $s+1$ values of λ so that $p-\lambda q$ is a sum of s $2s$-th powers. □

Sylvester [1870] recalled his discovery of this algorithm and its consequences.

> "I discovered and developed the whole theory of canonical binary forms for odd degrees, and, as far as yet made out, for even degrees too, at one evening sitting, with a decanter of port wine to sustain nature's flagging

energies, in a back office in Lincoln's Inn Fields. The work was done, and well done, but at the usual cost of racking thought — a brain on fire, and feet feeling, or feelingless, as if plunged in an ice-pail. *That night we slept no more.*"

Example 3.1. This example of Sylvester's algorithm will be used in Example 4.1. Let

$$\begin{aligned} p(x, y) &= 2x^3 + 3x^2y - 21xy^2 - 41y^3 \\ &= \tbinom{3}{0} \cdot 2\ x^3 + \tbinom{3}{1} \cdot 1\ x^2y + \tbinom{3}{2} \cdot (-7)\ xy^2 + \tbinom{3}{3} \cdot (-41)\ y^3. \end{aligned}$$

Since

$$\begin{pmatrix} 2 & 1 & -7 \\ 1 & -7 & -41 \end{pmatrix} \cdot \begin{pmatrix} 6 \\ -5 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

we have $h(x, y) = 6x^2 - 5xy + y^2 = (2x - y)(3x - y)$. It now follows that $p(x, y) = \lambda_1(x + 2y)^3 + \lambda_2(x + 3y)^3$, and a simple computation shows that $\lambda_1 = 5$, $\lambda_2 = -3$.

Lemma 3.1, when applied to Theorem 2.1, yields the following corollary.

Corollary 3.9. *A basis for $H_d(\mathbb{C}^n)$ is given by $\{(i \cdot)^d : i \in \mathcal{I}(n, d)\}$.*

This in turn gives a very weak version of the Alexander–Hirschowitz theorem:

Corollary 3.10. *A general form in $H_d(\mathbb{C}^n)$ is a sum of*

$$N(n, d - 1) = \frac{nd}{n + d - 1} \cdot \frac{1}{n} N(n, d)$$

d-th powers of linear forms.

Proof. Consider the sum

$$\sum_{l=1}^{N(n,d-1)} (t_{l,1}x_1 + \cdots + t_{l,n}x_n)^d,$$

and apply Corollary 2.3 with t_l specialized to $i_l \in \mathcal{I}(n, d - 1)$. Then J contains $x_k(i_l \cdot)^{d-1}$ for each k, l and hence $x_k H_{d-1}(\mathbb{C}^n) \subseteq J$ for each k, so $J = H_d(\mathbb{C}^n)$. $\square$

4. Examples of binary canonical forms and the proof of Theorem 1.6

This section is devoted to special cases of Theorem 1.6. First, in the special case $e_k = 1$, we give a constructive proof showing uniqueness, which gives a kind of interpolation between Sylvester's theorem and the representations of $H_d(\mathbb{C}^2)$ by (2-4) with a fixed basis consisting of d-th powers, as in Corollary 3.6.

Corollary 4.1. *Suppose $d \geq 1$, and $\{l_j(x, y) = \alpha_j x + \beta_j y\}$ is a fixed honest set of $m = d + 1 - 2r$ linear forms. Then a general binary d-ic form $p \in H_d(\mathbb{C}^2)$ can be written uniquely as*

$$p(x, y) = \sum_{j=1}^{m} t_j l_j(x, y)^d + \sum_{k=1}^{r} (t_{k1} x + t_{k2} y)^d \tag{4-1}$$

for suitable $t_{k1}, t_{k2} \in \mathbb{C}$.

Proof. Let

$$f(x, y) = \prod_{j=1}^{m} (\beta_j x - \alpha_j y).$$

Then $f(D)p$ has degree $d - m = 2r - 1$ and by Theorem 3.8 generally has a unique representation as a sum of r $(2r-1)$-st powers of linear forms, say

$$f(D)p = \sum_{k=1}^{r} (u_{k1} x + u_{k2} y)^{2r-1}. \tag{4-2}$$

Further, it is generally true that $f(u_{k1}, u_{k2}) \neq 0$. Let

$$q(x, y) = \frac{(2r-1)!}{d!} \sum_{k=1}^{r} \frac{(u_{k1} x + u_{k2} y)^d}{f(u_{k1}, u_{k2})}. \tag{4-3}$$

It follows from Theorem 3.3(iii), (4-2) and (4-3) that $f(D)p = f(D)q$. Since f has distinct factors, it then follows from Theorem 3.8 that there exist $t_j \in \mathbb{C}$ so that

$$p(x, y) - q(x, y) = \sum_{j=1}^{m} t_j (\alpha_j x + \beta_j y)^d.$$

Conversely, suppose p has two different representations:

$$\sum_{j=1}^{m} t_j l_j^d(x, y) + \sum_{k=1}^{r} (t_{k1} x + t_{k2} y)^d = \sum_{j=1}^{m} \tilde{t}_j l_j^d(x, y) + \sum_{k=1}^{r} (\tilde{t}_{k1} x + \tilde{t}_{k2} y)^d. \tag{4-4}$$

By combining the first sum on each side, (4-4) becomes a linear dependence with $m + 2r = d + 1$ summands, which by Lemma 3.7 must be trivial; thus, the representations in (4-4) are essentially the same. □

Example 4.1. Let $l_1(x, y) = x + y$ and $l_2(x, y) = -x + 3y$ and let

$$p(x, y) = -x^5 + 15x^4 y - 170x^3 y^2 + 390x^2 y^3 - 505x^2 y^3 + 483y^5.$$

In an application of the last proof, $f(x, y) = (x - y)(3x + y) = 3x^2 - 2xy - y^2$,

and

$$3\frac{\partial^2 p}{\partial x^2} - 2\frac{\partial^2 p}{\partial x \partial y} - \frac{\partial^2 p}{\partial y^2} = 160x^3 + 240x^2y - 1680xy^2 - 3280y^3.$$

Example 3.1 implies that this expression equals $400(x+2y)^3 - 240(x+3y)^3$. Since $f(1,2) = -5$ and $f(1,3) = -12$, it follows that

$$\begin{aligned} p(x,y) &= \frac{3!\cdot 400}{5!\cdot(-5)}(x+2y)^5 + \frac{3!\cdot(-240)}{5!\cdot(-12)}(x+3y)^5 + t_1(x+y)^5 + t_2(-x+3y)^5 \\ &= -4(x+2y)^5 + (x+3y)^5 + t_1(x+y)^5 + t_2(-x+3y)^5 \end{aligned}$$

and it can be readily be computed that $t_1 = \frac{7}{2}$ and $t_2 = \frac{3}{2}$.

If each $e_k = 2$ in Theorem 1.6 and m is as small as possible, then we obtain an analogue of Sylvester's theorem for forms of even degree.

Corollary 4.2. (i) *A general binary form of degree $d = 6s$ can be written as*

$$\lambda x^{6s} + \sum_{j=1}^{2s} (\alpha_j x^2 + \beta_j xy + \gamma_j y^2)^{3s} \quad \textit{for some } \lambda \in \mathbb{C}. \tag{4-5}$$

(ii) *A general binary form of degree $d = 6s+2$ can be written as*

$$\sum_{j=1}^{2s+1} (\alpha_j x^2 + \beta_j xy + \gamma_j y^2)^{3s+1}. \tag{4-6}$$

(iii) *A general binary form of degree $d = 6s+4$ can be written as*

$$\lambda_1 x^{6s+4} + \lambda_2 y^{6s+4} + \sum_{j=1}^{2s+1} (\alpha_j x^2 + \beta_j xy + \gamma_j y^2)^{3s+2} \quad \textit{for some } \lambda_i \in \mathbb{C}. \tag{4-7}$$

We have not been able to find an analogue to Sylvester's algorithm for determining the representations (4-5), (4-6), (4-7) in Corollary 4.2. In the linear case, $(\alpha x + \beta y)^d$ is killed by $\beta\,\partial/\partial x - \alpha\,\partial/\partial y$, and two operators of this shape commute. Although each $(\alpha x^2 + 2\beta xy + \gamma y^2)^d$ is killed by the nonconstant-coefficient $(\beta x + \gamma y)\partial/\partial x - (\alpha x + \beta y)\partial/\partial y$, two operators of this kind do not usually commute. The smallest constant-coefficient differential operator which kills $(\alpha x^2 + 2\beta xy + \gamma y^2)^d$ has degree $d+1$; the product of any two of these would kill every form of degree $2d$ and so provide no information.

Let us say that (1-8) is a *neat* canonical form if $m = 0$, and of *Sylvester-type* if it is neat and if $e_k = e$ for $1 \le k \le r$. Counting the numbers of neat and Sylvester-type canonical forms leads to some number theory. The first lemma is standard.

Lemma 4.3. *Given $0 < p/q \in \mathbb{Q}$ and $0 < n \in \mathbb{N}$, there exist only finitely many choices of $m_j \in \mathbb{Z}$, $0 < m_1 \le m_2 \cdots \le m_n$, such that $p/q = \sum_{j=1}^{n} 1/m_j$.*

Proof. If $n = 2$, then $p/q > 1/m_1 \geq p/(2q)$ implies that there are finitely many integral choices for m_1, each of which determines $m_2 = (p/q - 1/m_1)^{-1}$. Supposing the lemma valid for $n-1$, we have $p/q > 1/m_1 \geq p/(nq)$, and each choice of m_1 implies the equation $p/q - 1/m_1 = \sum_{j=2}^{n} 1/m_j$. This has finitely many solutions by the induction hypothesis. □

Theorem 4.4. *For a fixed value of r, there are only finitely many neat canonical forms (1-8) with r summands.*

Proof. Suppose $m = 0$ in Theorem 1.6. Write $d = e_k m_k$; then, by (1-7),

$$d + 1 = \sum_{k=1}^{r} \left(\frac{d}{m_k} + 1 \right),$$

which implies

$$1 = \sum_{k=1}^{r} \frac{1}{m_k} + \frac{r-1}{d} = \sum_{k=1}^{r} \frac{1}{m_k} + \sum_{l=1}^{r-1} \frac{1}{d}. \tag{4-8}$$

Now apply Lemma 4.3 with $p/q = 1$ and $n = 2r - 1$: there are only finitely many expressions of 1 as a sum of $2r-1$ unit fractions, of which only a subset satisfy the additional restrictions of (4-8). □

It is not hard to work out that, for $r = 2$, there are three neat canonical forms: $(d, e_1, e_2) = (3, 1, 1)$, $(4, 2, 1)$ and $(6, 3, 2)$. The first is Theorem 1.1(i) with $d = 3$, the second is Corollary 1.7 with $d = 4$ (see Theorem 4.6 below), and the third is Theorem 1.5. When $r = 3$, there are twenty-two neat canonical forms.

Let $s(d)$ denote the number of neat Sylvester-type canonical forms of degree d. Suppose $e_k = e$ for all k in one of these. Then $e \mid d$ and, by (1-7), $r(e+1) = d+1$, so $(e+1) \mid (d+1)$. Since $d \equiv 0 \pmod{e}$ and $d \equiv -1 \pmod{(e+1)}$, it follows from the Chinese remainder theorem that $d \equiv e \pmod{e(e+1)}$; that is, $d = e + ue(e+1)$, $u \geq 1$, so that $e < \sqrt{d}$.

Theorem 4.5. *Let $S(N) := \sum_{d=1}^{N} s(d)$. Then $S(N) = N + \mathbb{O}(N^{1/2})$, $\sup_d s(d) = \infty$.*

Proof. The generating function for the sequence $(s(d))$ is

$$\sum_{n=1}^{\infty} s(d) x^d = \sum_{e=1}^{\infty} \sum_{u=1}^{\infty} x^{e+ue(e+1)} = \sum_{e=1}^{\infty} \frac{x^{e^2+2e}}{1 - x^{e^2+e}} = \sum_{N=e}^{\infty} \left\lfloor \frac{N-e}{e^2+e} \right\rfloor X^N. \tag{4-9}$$

Let $T = \lfloor N^{1/2} \rfloor$. It follows from (4-9) that

$$S(N) = \sum_{n=1}^{N} s_n = \sum_{e=1}^{\infty} \left\lfloor \frac{N-e}{e^2+e} \right\rfloor = \sum_{e=1}^{T} \left\lfloor \frac{N-e}{e^2+e} \right\rfloor. \tag{4-10}$$

Thus, using the telescoping sum for $\sum \frac{1}{e(e+1)}$, (4-10) implies that

$$S(N) \le \sum_{e=1}^{T} \frac{N-e}{e^2+e} = N \sum_{e=1}^{T} \frac{1}{e^2+e} - \sum_{e=1}^{T} \frac{1}{e+1} \tag{4-11}$$
$$\le N\left(1 - \frac{1}{T+1}\right) - \log T + \mathbb{O}(1) = N - N^{1/2} + \mathbb{O}(\log N).$$

The lower bound is the same, minus T, so (4-11) implies that $S(N) = N + \mathbb{O}(N^{1/2})$.

Now, $s(d)$ counts the number of $e < d$ so that e divides d and $e+1$ divides $d+1$. If $d = 2^r - 1$, then $e+1 \mid 2^r$ implies that $e+1 = 2^t$ for some $t < r$. But $2^t - 1 \mid 2^r - 1$ if and only if $t \mid r$; hence $s(2^r - 1) = d(r) - 1$, where $d(n)$ denotes the divisor function. In particular, $s(2^{2^t} - 1) = t$, so the sequence $(s(d))$ is unbounded. More generally, if $e \mid d$ and $e+1 \mid d+1$, then $e \mid d^2 + 2d$ and $e+1 \mid d^2 + 2d + 1$, and since $e = d$ contributes to the count in $s(d^2 + 2d)$ but not in $s(d)$, $s(d^2 + 2d) \ge s(d) + 1$. □

Half of the neat Sylvester forms come from Theorem 1.1(i), another sixth come from Corollary 4.2(ii), etc. The smallest d for which $s(d) = 2$ is $d = 15$: $(e, r) = (1, 8), (3, 4)$, so a general binary form of degree 15 is a sum of eight linear forms to the 15th power, or four cubics to the 5th power. Mathematica computations show that the smallest d for which $s(d) = 3$ is $d = 99$: $(e, r) = (1, 50), (3, 25), (9, 10)$. For $d < 10^7$, the largest value of $s(d)$ is $s(7316000) = 12$. Note that $2^{2^{13}} - 1 = 2^{4096} - 1 \approx 1.04 \times 10^{1233}$, so the examples given in the proof are not likely to describe the fastest growth. We conjecture as well that $\{s(d)\}$ has an underlying distribution.

If the degree d is prime, then Corollary 4.1 accounts for all canonical forms in Theorem 1.6. The smallest d which is not covered by Corollary 4.1 is then $d = 4$, and there are two such cases, one of which is neat: $e_1 = 2$, $e_2 = 1$, $m = 0$ and $e_1 = 2$, $m = 2$. Both can be discussed constructively.

Theorem 4.6. *A general binary quartic $p \in H_4(\mathbb{C}^2)$ can be written as*

$$p(x, y) = (t_1 x^2 + t_2 xy + t_3 y^2)^2 + (t_4 x + t_5 y)^4 \tag{4-12}$$

in six different ways. Further, the set of possible values for $\{t_5/t_4\}$ is the image of the set $\{0, \infty, 1, -1, i, -i\}$ under a Möbius transformation.

Proof. By Theorem 2.4, if p is a general binary quartic, then there exist c_i, λ so that $p(c_1 x + c_2 y, c_3 x + c_4 y) = p_\lambda(x, y) := x^4 + 6\lambda x^2 y^2 + y^4$. If (4-12) holds for p_λ, then

$$1 = t_1^2 + t_4^4, \quad 0 = 2t_1 t_2 + 4t_4^3 t_5, \quad 6\lambda = 2t_1 t_3 + t_2^2 + 6t_4^2 t_5^2, \tag{4-13}$$
$$0 = 2t_2 t_3 + 4t_4 t_5^3, \quad 1 = t_3^2 + t_5^4.$$

First suppose that $t_4 = 0$. Then (4-13) implies that $1 = t_1^2$ and $0 = 2t_1t_2$, so $t_1 = 1$ (without loss of generality) and $t_2 = 0$. The remaining equations imply that $t_3 = 3\lambda$ and $t_5^4 = 1 - 9\lambda^2$. A similar argument works if $t_5 = 0$, giving two representations:

$$p_\lambda(x, y) = (x^2 + 3\lambda y^2)^2 + (1 - 9\lambda^2)y^4 = (3\lambda x^2 + y^2)^2 + (1 - 9\lambda^2)x^4. \tag{4-14}$$

Now suppose $t_4t_5 \neq 0$, so $t_1t_2t_3 \neq 0$ and we get successively

$$\frac{t_3}{t_1} = \frac{-2t_2t_3}{-2t_1t_2} = \frac{4t_4t_5^3}{4t_4^3t_5} = \frac{t_5^2}{t_4^2} \implies \frac{1 - t_3^2}{1 - t_1^2} = \frac{t_5^4}{t_4^4} = \frac{t_3^2}{t_1^2} \implies t_1^2 = t_3^2.$$

It follows that $t_5 = i^k t_4$ and $t_3 = (-1)^k t_1$, and (4-13) can be completely solved:

$$t_4^4 = 1 - t_1^2, \quad t_2 = 2i^k(t_1 - t_1^{-1}), \quad 2 + 6(-1)^k\lambda = 4t_1^{-2}.$$

After some massaging of the algebra, this gives four representations:

$$p_\lambda(x, y) = \left(\frac{(-1)^k 2}{3\lambda + (-1)^k}\right)\bigl(x^2 - i^{3k}(3\lambda - (-1)^k)xy + (-1)^k y^2\bigr)^2 + \left(\frac{3\lambda - (-1)^k}{3\lambda + (-1)^k}\right)(x + i^k y)^4, \quad k = 0, 1, 2, 3. \tag{4-15}$$

In order to find the six representations of p as (4-12), we start with the six representations of p_λ given in (4-14) and (4-15), in which $t_4x + t_5y$ is a multiple of one of the six linear forms x, y, $x + i^k y$. Apply the inverse of the map $(x, y) \mapsto (c_1x + c_2y, c_3x + c_4y)$, which takes $t_4x + t_5y$ to a multiple of $t_4(c_4x - c_2y) + t_5(-c_3x + c_1y)$: $t_5/t_4 \mapsto G(t_5/t_4)$, where $G(z) = (c_1z - c_2)/(c_4 - c_3z)$. □

Theorem 4.7. *Given two fixed nonproportional binary linear forms l_1, l_2, a general binary quartic in $H_4(\mathbb{C}^2)$ has two representations as*

$$p(x, y) = (t_1x^2 + t_2xy + t_3y^2)^2 + t_4l_1(x, y)^4 + t_5l_2(x, y)^4. \tag{4-16}$$

Proof. Given p, l_1, l_2, make an invertible linear change of variable taking $(l_1, l_2) \mapsto (x, y)$, and suppose $p(x, y) \mapsto q(x, y) = \sum_i a_i x^{4-i} y^i$. Then q has the shape (4-16) if and only if the coefficients of x^3y, x^2y^2, xy^3 in $(t_1x^2 + t_2xy + t_3y^2)^2$ and q agree. Thus, we seek to solve the system

$$a_1 = 2t_1t_2, \quad a_2 = 2t_1t_3 + t_2^2, \quad a_3 = 2t_2t_3. \tag{4-17}$$

But (4-17) implies $a_1t_2^2 - 2a_2t_1t_2 + 2a_3t_1^2 = 0$; hence, in general, there are exactly two values of β so that $t_2 = \beta t_1$; in each case, $t_1^2 = a_1/(2\beta)$. The two choices of sign for t_1 lead to the same square, and $t_3 = (a_1/a_3)t_1$, so (4-17) has these two solutions. □

In the case of Theorem 1.6 let $F(d; e_1, \ldots, e_r)$ denote the number of different representations that a general $p \in H_d(\mathbb{C}^2)$ has, by our convention. We present in Table 1

d	$e_1, \dots, e_r$	m	$F(d; e)$	Source
any	$1^{\lfloor (d+1)/2 \rfloor}$	0 or 1	1	Theorem 1.1
any	1^r	$d+1-2r$	1	Corollary 4.1
4	2, 1	0	6	Theorem 4.6
4	2	2	2	Theorem 4.7
6	3, 2	0	40	[Várilly-Alvarado 2008; 2011]
6	2, 1^2	0	22	Experiment
6	3, 1	1	14	Experiment
6	2^2	1	9	Experiment
6	2, 1	2	12	Experiment
6	3	3	5	Experiment
6	2	4	5	Experiment
8	2, 1^3	0	62	Experiment
10	2, 1^4	0	147	Experiment
12	2, 1^5	0	308	Experiment
$2s$	2, 1^{s-1}	0	$2\binom{s+3}{5}+\binom{s+2}{3}$	Conjecture

Table 1. Proved and conjectural values of $F(d; e)$.

a complete list of proved or conjectural values when $d \le 6$, reflecting numerical experiments on Mathematica. (Recall that, if d is prime, then Corollary 4.1 presents all possible canonical forms of this type.) The conjectural value of $F(2s; 2, 1^{s-1})$ is suggested by the given data for $2 \le s \le 6$ and [OEIS 2013, A081282].

Várilly-Alvarado [2008; 2011] constructs explicitly all 240 representations of $x^6 + y^6$ as $f^2 + g^3$; he considers forms multiplied by roots of unity as different, which explains the appearance of $240/(2 \cdot 3)$ in the table above. This is also proved to be the number of representations for a general sextic.

To describe the experiments for $F(2s; 2, 1^{s-1})$ more precisely, we generate a form

$$p(x, y) = \sum_{k=0}^{2s} \binom{2s}{k} a_k x^{2s-k} y^k,$$

where $a_k = t + iu$ for random integers t, u in $[-100, 100]$. In case $s = 1$, we assume a change of variables so that the fixed linear forms are x^d or y^d; for $s \ge 2$ we choose additional linear forms with random coefficients. Let $h(x, y) = Ux^2 + Vxy + Wy^2$ for variables (U, V, W) and let $q(x, y) = p(x, y) - h^s(x, y)$, and apply Sylvester's algorithm to q. That is, we construct the $(s+2) \times s$ matrix $A_{s-1}(q)$, with polynomial entries in (U, V, W) of degree s and require that it have rank less than s. This is done by counting the number of (U, V, W) which are common zeros of all $s \times s$ minors. This number is divided by s to account for $h^s = (\zeta_s^k h)^s$. As a back of the envelope calculation, one might take the first $s-1$ rows of A_{s-1} and use the cofactors to compute a nontrivial null-vector. Ignoring possible cancellation, the components would be polynomials of degree $s(s-1)$ in (U, V, W). Taking the dot product with the last three rows of A_{s-1} gives three polynomials of degree s^2. Ignoring cancellations and multiplicity, there should be $(s^2)^3$ common zeros, and dividing by s gives an upper bound for $F(2s; 2, 1^{s-1})$ of s^5. The conjectural value is asymptotically $\frac{1}{60}s^5$, which shows the same order of growth.

5. Quadratic forms and the proof of Theorem 1.2

We begin this section with a constructive proof of Theorem 1.2 which will serve as a template for constructive proofs involving cubic forms.

Constructive Proof of Theorem 1.2. Suppose $p \in H_2(\mathbb{C}^n)$, and, specifically,

$$p(x_1, \dots, x_n) = \sum_{i=1}^{n} a_{ii}x_i^2 + 2 \sum_{1 \le i < j \le n} a_{ij}x_i x_j.$$

Then $\partial p/\partial x_1 = 2\sum_{j=1}^{n} a_{1j}x_j$. Since $a_{11} \neq 0$ in general, we can define

$$q(x_1, \dots, x_n) = p(x_1, \dots, x_n) - \frac{1}{a_{11}}\left(\sum_{j=1}^{n} a_{1j}x_j\right)^2.$$

Observe that $\partial q/\partial x_1 = 0$, so $q = q(x_2, \dots, x_n)$. Iterating this argument gives the construction. There is only one linear form $\pm l$ so that $\partial p/\partial x_1 = 2l\,\partial l/\partial x_1$, so the representation is unique. □

Constant-counting for sums of squares is complicated by the action of the orthogonal group on a sum of t squares. If $M \in \mathrm{Mat}_t(\mathbb{C})$ and $MM^t = I$, then

$$\sum_{i=1}^{t} f_i^2 = \sum_{i=1}^{t}\left(\sum_{j=1}^{t} m_{ij}f_j\right)^2.$$

When $t = 2$, choose $\theta \in \mathbb{C}$ and let $e^{i\theta} = \cos\theta + i\sin\theta := (u, v)$, so that

$$f^2 + g^2 = (uf - vg)^2 + (vf + ug)^2. \tag{5-1}$$

This means that we may safely remove one monomial from one of the summands.

Theorem 5.1. *A general binary form $p \in H_{2s}(\mathbb{C}^2)$ can be written as*

$$\left(\sum_{k=0}^{s} t_k x^{s-k}y^k\right)^2 + \left(\sum_{k=1}^{s} t_{s+k}x^{s-k}y^k\right)^2 \tag{5-2}$$

in $\binom{2s-1}{s}$ different ways.

Proof. The nonconstructive proof is a simple application of Corollary 2.3. Writing (5-2) as f^2+g^2 gives the partials with respect to the t_j as

$$\{2x^{s-k}y^k f, 0 \le k \le s\} \cup \{2x^{s-k}y^k g, 1 \le k \le s\};$$

specializing to $f = x^s$ and $g = y^s$ above gives all monomials in $H_{2s}(\mathbb{C}^2)$.

The more obvious expression

$$p(x,y) = f^2(x,y) + g^2(x,y), \quad g,h \in H_s(\mathbb{C}^2) \tag{5-3}$$

is not a canonical form, because $2(s+1) > 2s+1$. However, every sum of two squares can be formally factored, and these behave nicely with respect to (5-1):

$$f^2+g^2 = (f+ig)(f-ig) \quad \Longleftrightarrow$$
$$(uf+vg)^2+(vf-ug)^2 = \big(e^{i\theta}(f+ig)\big)\big(e^{-i\theta}(f-ig)\big).$$

Suppose $p(1,0) = a_0 \ne 0$ (true for general p) and (5-3) holds, where $f(1,0) = \rho$ and $g(1,0) = \tau$. Then $\rho^2+\tau^2 = a_0$, so that $\tau/\rho \ne \pm i$ and the coefficient of x^s in $vf+ug$ will be $v\rho+u\tau = \sin\theta\rho+\cos\theta\tau$, which is zero exactly when $\tan\theta = -\tau/\rho$. Thus, for precisely one value of $\tan\theta$, the right-hand side of (5-1) will be in the form (5-2). This determines a pair $(\pm u, \pm v)$; however, the squares in (5-2) will be the same.

In other words, each distinct factorization of p (up to multiple) as a product of two s-ic forms, when combined with the orthogonal action of (5-1), yields exactly one representation as (5-2). A general $p \in H_{2s}(\mathbb{C}^2)$ is a product of $2s$ distinct linear factors; these can be organized into an unordered pair of products of s distinct linear factors in $\frac{1}{2}\binom{2s}{s} = \binom{2s-1}{s}$ ways. □

The "lost" degree of freedom in a sum of squares never arises in Theorem 1.6 because $2(d/2+1) > d+1$. The missing monomial x^s in the second summand of (5-2) may be replaced by any specified monomial $x^{s-k_0}y^{k_0}$ by a similar argument.

Another classical result is that a general ternary quartic is a sum of three squares of quadratic forms, generally in 63 different ways up to the action of the orthogonal group (see [Powers et al. 2004]). Hilbert [1888] proved that *every* positive semidefinite $p \in H_4(\mathbb{R}^3)$ is a sum of three squares from $H_2(\mathbb{R}^3)$. He then showed that there exist psd forms in $H_6(\mathbb{R}^3)$ and $H_4(\mathbb{R}^4)$ which are not sums of squares in $H_3(\mathbb{R}^3)$ and $H_2(\mathbb{R}^4)$, respectively, which ultimately led to his 17th problem. (See [Reichstein 1987] for much more on this subject.)

A constructive discussion of Hilbert's theorem on $p \in H_4(\mathbb{R}^3)$ has recently been given in [Powers and Reznick 2000; Powers et al. 2004; Pfister and Scheiderer 2012; Plaumann et al. 2011]. A nonconstructive proof (without the count) can easily be given.

Theorem 5.2. *A general ternary quartic $p \in H_4(\mathbb{C}^3)$ can be written as $p = q_1^2 + q_2^2 + q_3^2$, where $q_j \in H_2(\mathbb{C}^3)$.*

Proof. We take q_i so that the monomial x^2 only appears in q_1 and the monomial y^2 only appears in q_1 and q_2, and so the number of coefficients in the q_j is $6+5+4=15$. Taking the partials where $(q_1, q_2, q_3) = (x^2, y^2, z^2)$ shows that J contains $2x^2\{x^2, y^2, z^2, xy, xz, yz\}$, $2y^2\{y^2, z^2, xy, xz, yz\}$ and $2z^2\{z^2, xy, xz, yz\}$, and so is equal to $H_4(\mathbb{C}^3)$. $\square$

Since $3\binom{m+2}{2} - 3 < \binom{2m+1}{2}$ for $m \geq 3$, this result does not generalize to ternary forms of higher even degree.

The situation is somewhat simpler over $\mathbb{R}$. A real version of Theorem 5.1 appears in [Reznick 2000]. If p is real and positive definite and $p = f^2 + g^2$, where f and g are also real, then the factors of p consist of s conjugate pairs. In the factorization $p = (f + ig)(f - ig)$, the pairs must be split between the conjugate factors, and if p has distinct factors, this can be done in 2^{s-1} different ways. A real generalization of Theorem 5.2 appears in [Choi et al. 1995, Corollary 2.12]. Suppose a real psd form $p \in H_{2s}(\mathbb{R}^n)$ is a sum of t squares and $x^{\beta_i} \in H_s(\mathbb{R}^n)$, $1 \leq i \leq t$, is given. Then there is a representation $p = \sum_{j=1}^{t} g_j^2$, in which x^{β_i} does not occur in g_j for $j > i$. This argument can also be applied to a *general* sum of t squares over $\mathbb{C}$, but it no longer applies to all forms. For example, if $xy = (ax + by)^2 + (cx + dy)^2$, then $abcd \neq 0$.

6. Cubic forms and proofs of Theorems 1.3 and 1.4

In this section, we present three representations for forms in $H_3(\mathbb{C}^n)$ as a sum of cubes of linear forms. The first two are canonical; the third isn't, but it represents *all* cubics, not just general cubics.

We begin with Theorem 1.3, which first appeared in [Reichstein 1987]. At the time of this writing, that paper has had no citations in MathSciNet. (It was discussed in [Reznick 1992b] and, from there, in [Comon and Mourrain 1996]. The former was never submitted for publication and the latter appeared in an unindexed journal.) The original presentation and proof in [Reichstein 1987] were given for trilinear forms (see Section 2); the theorem is applied to cubic forms there mainly in the examples.

By iterating (1-5), we obtain a canonical form for $p \in H_3(\mathbb{C}^n)$; see [Reichstein 1987, p. 98].

Corollary 6.1. *A general n-ary cubic $p \in H_3(\mathbb{C}^n)$ can be written uniquely as*

$$p(x_1, \ldots, x_n) = \sum_{m=0}^{\lfloor (n-1)/2 \rfloor} \sum_{k=1}^{n-2m} \left(t_{m,1+2m}^{\{k\}} x_{1+2m} + \cdots + t_{m,n}^{\{k\}} x_n\right)^3 \tag{6-1}$$

for some $t_{m,j}^{\{k\}} \in \mathbb{C}$.

This gives p as a sum of $n+(n-2)+\cdots = \lfloor (n+1)^2/4 \rfloor$ cubes. Recall that, by Alexander–Hirschowitz, for $n \neq 5$, a general cubic form in n variables can be written as a sum of $\lceil (n+1)(n+2)/6 \rceil$ cubes. Thus (6-1) is a canonical form which represents a general cubic as a sum of about 50% more cubes than the true minimum; this is due to the large number of linear forms with restricted sets of variables.

Reichstein's proof of Theorem 1.3 requires the well-known "generalized eigenvalue problem" for pairs of symmetric matrices, as interpreted for quadratic forms: if a general pair of quadratic forms $f,\ g \in H_2(\mathbb{C}^n)$ is given, then there exist n linearly independent forms $L_i(x) = \sum_{j=1}^{n} \alpha_{ij} x_j$ and $c_i \in \mathbb{C}$ so that

$$f = \sum_{i=1}^{n} L_i^2, \quad g = \sum_{i=1}^{n} c_i L_i^2. \tag{6-2}$$

If M_f, M_g are the matrices associated to f, g, then the c_i are the n roots of the determinantal equation $\det(M_g - \lambda M_f) = 0$, which are generally distinct, so the L_i are uniquely determined up to multiple. We may also assume that the coefficients α_{ij} of the linear forms are generally nonzero; cf. Corollary 6.3.

Proof of Theorem 1.3. For general $p \in H_3(\mathbb{C}^n)$, we simultaneously diagonalize $f = \partial p/\partial x_1$ and $g = \partial p/\partial x_2$ as in (6-2). Since mixed partials are equal,

$$\frac{\partial f}{\partial x_2} = \frac{\partial g}{\partial x_1} = \sum_{i=1}^{n} 2\alpha_{i2} L_i = \sum_{i=1}^{n} 2 c_i \alpha_{i1} L_i, \tag{6-3}$$

and since the L_i are linearly independent, (6-3) implies that $\alpha_{i2} = c_i \alpha_{i1}$.

It is generally true that $\alpha_{i1} \neq 0$. Let

$$q(x_1, \dots, x_n) = p(x_1, \dots, x_n) - \sum_{i=1}^{n} \frac{1}{3\alpha_{i1}} L_i^3.$$

It follows that

$$\frac{\partial q}{\partial x_1} = \frac{\partial p}{\partial x_1} - \sum_{i=1}^{n} \frac{3\alpha_{i1}}{3\alpha_{i1}} L_i^2 = \frac{\partial p}{\partial x_1} - \sum_{i=1}^{n} L_i^2 = 0,$$
$$\frac{\partial q}{\partial x_2} = \frac{\partial p}{\partial x_2} - \sum_{i=1}^{n} \frac{3\alpha_{i2}}{3\alpha_{i1}} L_i^2 = \frac{\partial p}{\partial x_2} - \sum_{i=1}^{n} c_i L_i^2 = 0.$$

Since $\partial q/\partial x_1 = \partial q/\partial x_2 = 0$, we have $q = q(x_3, \dots, x_n)$.

For uniqueness, suppose (1-5) holds and $l_k(x_1, \dots, x_n) = \sum_j \beta_{kj} x_j$. Then

$$\frac{\partial p}{\partial x_1} = \sum_{k=1}^{n} 3\beta_{k1} l_k^2, \quad \frac{\partial p}{\partial x_2} = \sum_{k=1}^{n} 3\beta_{k2} l_k^2.$$

Thus, after a scaling, $\partial p/\partial x_1$ and $\partial p/\partial x_2$ have already been simultaneously diagonalized (as in (6-2)), and the l_k are, up to multiples, a rearrangement of the L_k. □

We now give a constructive proof of Theorem 1.4, which gives a different canonical form for $H_3(\mathbb{C}^n)$ requiring even more cubes.

Proof of Theorem 1.4. The constant-counting makes this a potential canonical form: the variables are $t_{\{i,j\},k}$ with $1 \le i \le j \le k \le n$, and there are $\binom{n+2}{3} = N(n,3)$ such triples (i, j, k). Given $p \in H_3(\mathbb{C}^n)$, $\partial p/\partial x_n$ is a quadratic form, so we can generally complete the square by Theorem 1.2:

$$\frac{\partial p}{\partial x_n} = \sum_{j=1}^{n} (t_{jj}x_j + \cdots + t_{jn}x_n)^2.$$

Then $t_{jn} \neq 0$ for general p and if we let

$$q(x_1,\dots,x_n) = p(x_1,\dots,x_n) - \sum_{j=1}^{n} \frac{1}{3t_{jn}} (t_{jj}x_j + \cdots + t_{jn}x_n)^3,$$

then $\partial q/\partial x_n = 0$, so $q = q(x_1,\dots,x_{n-1})$. Iterate this construction to get (1-6).

Uniqueness follows by working backwards. If (1-6) holds for a cubic p, then it gives $\partial p/\partial x_n$ in its (unique) upper-triangular diagonalization. This can be integrated with respect to x_n and subtracted from p, giving a cubic $q(x_1,\dots,x_{n-1})$. Again, iterate. □

It is not hard to give nonconstructive proofs of Theorems 1.3 and 1.4 using Corollary 2.3. These are left for the reader.

We first presented this next construction in [Reznick 1992b]; an outline of the proof can be found in [Comon and Mourrain 1996]. This is not a canonical form, but is included here because it gives an absolute upper bound for the length of cubic forms.

Theorem 6.2. *If $p \in H_3(\mathbb{C}^n)$, then there exists an invertible linear change of variables $y_j = \sum \lambda_{jk}x_k$ and n linear forms l_j so that, for some $q \in H_3(\mathbb{C}^{n-1})$,*

(6-4)
$$p(x_1,\dots,x_n) = \sum_{j=1}^{n} l_j^3(x_1,\dots,x_n) + q(y_2,\dots,y_n).$$

Thus every cubic in n variables is a sum of at most $\binom{n+1}{2}$ cubes of linear forms.

Proof. Define linear forms $l_{j,m}(y)$ for $1 \le j \le m+1$ by

(6-5)
$$l_{j,m}(y_1,\dots,y_n) = y_j + \alpha \sum_{j=1}^{m} y_j, \quad 1 \le j \le m$$

and

$$l_{m+1,m}(y_1,\dots,y_n) = -(1+m\alpha)\sum_{j=1}^{m} y_j, \quad \alpha = \frac{-(m+1)+\sqrt{m+1}}{m(m+1)}. \tag{6-6}$$

Then it can be easily checked that

$$\sum_{j=1}^{m+1} l_{j,m}(y) = 0 \quad \text{and} \quad \sum_{j=1}^{m+1} l_{j,m}^2(y) = \sum_{k=1}^{m} y_k^2. \tag{6-7}$$

Suppose $0 \neq p \in H_3(\mathbb{C}^n)$. Use Biermann's theorem to find a point u where $p(u) \neq 0$, and, after an invertible linear change of variables, taking $\{x_j\} \mapsto \{u_j\}$, we may assume that $p(1,0,\dots,0) = 1$ and so

$$p = u_1^3 + 3h_1(u_2,\dots,u_n)u_1^2 + 3h_2(u_2,\dots,u_n)u_1 + h_3(u_2,\dots,u_n), \tag{6-8}$$

where $\deg h_j = j$. Now let $u_1 = y_1 - h_1(u_2,\dots,u_n)$ to clear the quadratic term, so

$$p = y_1^3 + 3y_1\tilde{h}_2(u_2,\dots,u_n) + \tilde{h}_3(u_2,\dots,u_n), \tag{6-9}$$

where again $\deg \tilde{h}_j = j$. Diagonalize $\tilde{h}_2(u_2,\dots,u_n)$ as a quadratic form into $y_2^2 + \dots + y_r^2$, where $r \le n$, and make the accompanying change of variables. We now have

$$p = y_1^3 + 3y_1(y_2^2 + \dots + y_r^2) + k_3(y_2,\dots,y_n), \quad r \le n, \tag{6-10}$$

where $\deg k_3 = 3$. Finally, using (6-5) and (6-7), we construct g, a sum of $r \le n$ cubes:

$$\begin{aligned} g(y_1,\dots,y_n) &:= \frac{1}{r}\sum_{j=1}^{r}\bigl(y_1 + \sqrt{r}\cdot l_{j,r-1}(y_2,\dots,y_r)\bigr)^3 \\ &= \frac{1}{r}\sum_{j=1}^{r} y_1^3 + \frac{3}{\sqrt{r}}\sum_{j=1}^{r} y_1^2 l_{j,r-1} + 3\sum_{j=1}^{r} y_1 l_{j,r-1}^2 + \sqrt{r}\sum_{j=1}^{r} l_{j,r-1}^3 \\ &= y_1^3 + 3y_1(y_2^2 + \dots + y_r^2) + \sqrt{r}\sum_{j=1}^{r} l_{j,r-1}^3(y_2,\dots,y_r). \end{aligned} \tag{6-11}$$

Then

$$q := p - g$$

is a cubic form in $(y_2,\dots,y_n)$ as in (6-4). Iteration of this argument shows that any cubic $p \in H_3(\mathbb{C}^n)$ is a sum of at most $n(n+1)/2$ cubes. $\square$

Equation (1-5) can be extended to a canonical form for quartics as a sum of fourth powers of linear forms. Note that x_n appears in each summand of (6-1), with, generally, a nonzero coefficient.

Corollary 6.3. *For general $p \in H_4(\mathbb{C}^n)$, there exist $l_k \in H_1(\mathbb{C}^n)$ and $q \in H_4(\mathbb{C}^{n-1})$ so that, with $a(n) = \lfloor (n+1)^2/4 \rfloor$,*

$$p(x_1, \dots, x_n) = \sum_{k=1}^{a(n)} l_k(x_1, \dots, x_n)^4 + q(x_1, \dots, x_{n-1}).$$

As a consequence, a general $p \in H_4(\mathbb{C}^n)$ can be written as

$$p(x_1, \dots, x_n) = \sum_{m=0}^{\lfloor (n-1)/2 \rfloor} \sum_{r=1+2m}^{n} \sum_{k=1}^{r-2m} \bigl(t_{m,r,1+2m}^{\{k\}} x_{1+2m} + \cdots + t_{m,r,r}^{\{k\}} x_r\bigr)^4.$$

Proof. By Theorem 1.3 and (6-1), for general $p \in H_4(\mathbb{C}^n)$, we can write

$$\text{(6-12)} \qquad \begin{aligned} \frac{\partial p}{\partial x_n} &= \sum_{m=0}^{\lfloor (n-1)/2 \rfloor} \sum_{k=1}^{n-2m} \bigl(t_{m,1+2m}^{\{k\}} x_{1+2m} + \cdots + t_{m,n}^{\{k\}} x_n\bigr)^3 \\ &=: \sum_{m=0}^{\lfloor (n-1)/2 \rfloor} \sum_{k=1}^{n-2m} (l_m^{\{k\}}(x))^3. \end{aligned}$$

As before, if $q = p - \sum_{k,m} 1/(4t_{m,n}^{\{k\}}) l_{k,m}^4$, then $\partial q/\partial x_n = 0$, so $q = q(x_1, \dots, x_{n-1})$. Repeat as before. There are $N(n,3)$ coefficients in (6-12), and, since $N(n,3) + N(n-1,4) = N(n,4)$, the count is correct for a canonical form. □

Note that there is no variable which appears in each linear form in (6-12), so the argument can't be extended to quintics. For the same reason, Theorem 1.4 does not extend to quartics. By combining Theorems 1.3 and 6.3, we obtain canonical forms as a sum of powers of linear forms in the four exceptional cases of Alexander–Hirschowitz, of course at the expense of the number of summands. With regard to ternary quartics and Theorem 1.10, Corollary 6.3 becomes the following canonical form for $H_4(\mathbb{C}^3)$ as a sum of seven fourth powers:

$$\sum_{k=1}^{3} (t_{k1}x_1 + t_{k2}x_2 + t_{k3}x_3)^4 + t_{10}x_3^4 + \sum_{l=1}^{2} (u_{l1}x_1 + u_{l2}x_2)^4 + u_5 x_1^4.$$

There is an arithmetic obstruction to a "Reichstein-type" canonical form for quartics, that is, one in which each linear form is allowed to involve each variable. If

$$\text{(6-13)} \qquad p(x_1, \dots, x_n) = \sum_{k=1}^{r} (\alpha_{k1}x_1 + \cdots + \alpha_{kn}x_n)^4 + q(x_1, \dots, x_m)$$

were a canonical form for some n, then we would have $N(n,4) = rn + N(m,4)$. However, for $n = 12$, there does not exist $m < 12$ so that $12 \mid \binom{15}{4} - \binom{m+3}{4}$, so no such canonical form can exist. More generally, let

$$A_d = \left\{n \,:\, n \nmid \binom{n+d-1}{d} - \binom{m+d-1}{d} \text{ for } 0 \le m < n\right\} \tag{6-14}$$

denote the set of n for which this argument rules out Reichstein-type canonical forms. We present without proof a number of results about A_d. Note that there is no obstacle for (6-13) in prime degree, such as $d = 2, 3$.

Proposition 6.4. (i) *If* $3 \nmid k$, *then* $n = 2^{2k} \cdot 3 \in A_4$.

(ii) *If* $p \equiv 1 \pmod{144}$ *is prime*, *then* $12p \in A_4$.

(iii) *If* p *is prime*, *then* $p \mid \binom{n+p-1}{p} - \binom{n}{p}$; *hence* $A_p = \varnothing$ *for prime* p.

(iv) *The smallest elements of* A_6, A_8, A_{10}, A_{12}, A_{14} *and* A_{15} *are* 10, 1792, 6, 242, 338 *and* 273, *respectively. If* A_9 *or* A_{16} *are nonempty*, *then their smallest elements are at least* 10^5.

7. Subspace canonical forms and the proof of Theorem 1.11

One natural generalization of the definition of canonical forms is to consider maps $F : X \mapsto H_d(\mathbb{C}^n)$ where $X \subset \mathbb{C}^M$ is an $N(n, d)$-dimensional subspace of $\mathbb{C}^M$. (Similar ideas can be found in [Wakeford 1920], though his approach is different from ours.) These can be analyzed in the simplest nontrivial case: $M = 4$, $N(2, 2) = 3$.

Proof of Theorem 1.11. Assume that some $c_j \ne 0$. Without loss of generality, we may assume that $c_4 \ne 0$ and divide through by c_4 so that the equation is $t_4 = a_1t_1 + a_2t_2 + a_3t_3$, where $a_i = -c_i/c_4$ for $i = 1, 2, 3$. Then (1-13) can be parametrized as a map from $\mathbb{C}^3 \mapsto H_2(\mathbb{C}^2)$ as

$$F(t; x) = (t_1x + t_2y)^2 + \big(t_3x + (a_1t_1 + a_2t_2 + a_3t_3)y\big)^2. \tag{7-1}$$

The partials with respect to the t_j are

$$\begin{gathered} 2x(t_1x + t_2y) + 2a_1y\big(t_3x + (a_1t_1 + a_2t_2 + a_3t_3)y\big), \\ 2y(t_1x + t_2y) + 2a_2y\big(t_3x + (a_1t_1 + a_2t_2 + a_3t_3)y\big), \\ 2(x + a_3y)\big(t_3x + (a_1t_1 + a_2t_2 + a_3t_3)y\big). \end{gathered} \tag{7-2}$$

Now, (7-1) is a canonical form if and only if there exists a choice of t_i so that the three quadratics in (7-2) span $H_2(\mathbb{C}^2)$. A computation shows that the determinant of the forms in (7-2) with respect to the basis $\{x^2, xy, y^2\}$ is the cubic

$$-8\big((a_1a_2 - a_3)t_1 + (1 + a_2^2)t_2 + (a_2a_3 + a_1)t_3\big)(a_1t_1^2 + a_2t_1t_2 + a_3t_1t_3 - t_2t_3). \tag{7-3}$$

The second factor in (7-3) always has the term $-t_2t_3$ and so never vanishes; hence this determinant is not identically zero (and (7-1) is a canonical form), unless

$$a_1a_2 - a_3 = 1 + a_2^2 = a_2a_3 + a_1 = 0. \tag{7-4}$$

In the exceptional case where (7-4) holds, then $a_2 = \epsilon$, where $\epsilon = \pm i$, and $a_3 = \epsilon a_1$. Evaluating (7-1) at $(x, y) = (a_1, \epsilon)$ yields

$$(a_1t_1+\epsilon t_2)^2+(a_1t_3+\epsilon a_1t_1+\epsilon^2t_2+\epsilon^2a_1t_3)^2$$
$$=(a_1t_1+\epsilon t_2)^2+\big((1+\epsilon^2)a_1t_3+\epsilon a_1t_1+\epsilon^2t_2\big)^2=(a_1t_1+\epsilon t_2)^2+\epsilon^2(a_1t_1+\epsilon t_2)^2=0,$$

as claimed. □

It would be interesting to know how Theorem 1.11 generalizes to higher degrees.

Conjecture 1.12 is true for degree 2 by Theorem 1.11. We have verified it for even degrees up to eight by Corollary 2.3 applied to random choices for α_j, β_j in (1-14). We hold some hope that generalizations such as Conjecture 1.12 will have applications in more than two variables as well.

Acknowledgements

The greatest debt of the author is due Richard Ehrenborg and Gian-Carlo Rota for writing [Ehrenborg and Rota 1993]. Thanks to Dave Anderson and Julianna Tymoczko for organizing the Special Session on Geometric Commutative Algebra and Applications at the March 2011 AMS Sectional Meeting in Iowa City, to Lek-Heng Lim for organizing the Minisymposium on Tensor Rank at the October 2011 SIAM Conference on Applied Algebraic Geometry in Raleigh and to Eugene Mukhin for organizing ALGECOM5 in Indianapolis in October 2011. Invitations to speak at these conferences provided an opportunity to present preliminary versions of this material.

I am indebted to Tony Várilly-Alvarado for extremely helpful correspondence. I also want to thank T. Y. Lam, Hal Schenck, Bernd Sturmfels and Doron Zeilberger for their assistance. Special thanks go to the Center for Advanced Study at UIUC, where the author was an Associate in the Fall 2011 semester, and thereby free of teaching responsibilities. Finally, I want to thank the referee for a careful reading of the manuscript and for making many useful suggestions.

References

[Alexander and Hirschowitz 1995] J. Alexander and A. Hirschowitz, "Polynomial interpolation in several variables", *J. Algebraic Geom.* **4**:2 (1995), 201–222. MR 96f:14065 Zbl 0829.14002

[Babbage 1976] D. W. Babbage, ""Canonical forms for general polynomials" (J. London Math. Soc. (2) **10**:3 (1975), 271–280)—a note on a paper by Stefano Guazzone", *J. London Math. Soc.* (2) **13**:3 (1976), 564. MR 54 #5264 Zbl 0338.13007

[Carlini et al. 2012] E. Carlini, M. V. Catalisano, and A. V. Geramita, "The solution to the Waring problem for monomials and the sum of coprime monomials", *J. Algebra* **370** (2012), 5–14. MR 2966824 Zbl 06162665

[Choi et al. 1995] M. D. Choi, T. Y. Lam, and B. Reznick, "Sums of squares of real polynomials", pp. 103–126 in *K-theory and algebraic geometry: connections with quadratic forms and division* *algebras, II* (Santa Barbara, CA, 1992), edited by B. Jacob and A. Rosenberg, Proc. Sympos. Pure Math. **58**, Amer. Math. Soc., Providence, RI, 1995. MR 96f:11058 Zbl 0821.11028

[Clebsch 1861] A. Clebsch, "Ueber Curven vierter Ordnung", *J. Reine Angew. Math.* **59** (1861), 125–145. Zbl 059.1559cj

[Comon and Mourrain 1996] P. Comon and B. Mourrain, "Decomposition of quantics in sums of powers of linear forms", *Signal Process.* **53**:2–3 (1996), 93–107. Zbl 0875.94079

[Cox et al. 2007] D. Cox, J. Little, and D. O'Shea, *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*, 3rd ed., Springer, New York, 2007. MR 2007h:13036 Zbl 1118.13001

[Ehrenborg and Rota 1993] R. Ehrenborg and G.-C. Rota, "Apolarity and canonical forms for homogeneous polynomials", *Eur. J. Combin.* **14**:3 (1993), 157–181. MR 94e:15062 Zbl 0784.05019

[Elliott 1913] E. B. Elliott, *An introduction to algebra of quantics*, 2nd ed., Clarendon, Oxford, 1913. Reprinted by Chelsea, New York, 1964. JFM 44.0155.05

[Fisher 1967] C. S. Fisher, "The last invariant theorists: a sociological study of the collective biographies of mathematical specialists", *Eur. J. Sociol.* **8**:2 (1967), 216–244.

[Fröberg et al. 2012] R. Fröberg, G. Ottaviani, and B. Shapiro, "On the Waring problem for polynomial rings", *Proc. Natl. Acad. Sci. USA* **109**:15 (2012), 5600–5602. MR 2935563 Zbl 06133962

[Geramita 1996] A. V. Geramita, "Inverse systems of fat points: Waring's problem, secant varieties of Veronese varieties and parameter spaces for Gorenstein ideals", pp. 2–114 in *Curves Seminar at Queen's, 10* (Kingston, ON, 1995), edited by A. V. Geramita, Queen's Papers in Pure and Appl. Math. **102**, Queen's Univ., Kingston, ON, 1996. MR 97h:13012 Zbl 0864.14031

[Grace 1918] J. H. Grace, "Nachruf auf Edward Kingsley Wakeford (1894–1916)", *Proc. Lond. Math. Soc.* (2) **16** (1918), 54–57. JFM 46.0035.22

[Guazzone 1975] S. Guazzone, "Canonical forms for general polynomials", *J. London Math. Soc.* (2) **10** (1975), 271–280. MR 52 #3183 Zbl 0307.14014

[Gundelfinger 1887] S. Gundelfinger, "Zur Theorie der binären Formen", *J. Reine Angew. Math.* **100** (1887), 413–424. JFM 19.0109.01

[Hilbert 1888] D. Hilbert, "Ueber die Darstellung definiter Formen als Summe von Formenquadraten", *Math. Ann.* **32**:3 (1888), 342–350. MR 1510517 JFM 20.0198.02

[Iarrobino and Kanev 1999] A. Iarrobino and V. Kanev, *Power sums, Gorenstein algebras, and determinantal loci*, Lecture Notes in Math. **1721**, Springer, Berlin, 1999. MR 2001d:14056 Zbl 0942.14026

[Kung 1986] J. P. S. Kung, "Gundelfinger's theorem on binary forms", *Stud. Appl. Math.* **75**:2 (1986), 163–169. MR 87m:11020 Zbl 0607.10016

[Kung 1987] J. P. S. Kung, "Canonical forms for binary forms of even degree", pp. 52–61 in *Invariant theory* (West Chester, PA, 1985), edited by S. S. Koh, Lecture Notes in Math. **1278**, Springer, Berlin, 1987. MR 89h:15037 Zbl 0647.15019

[Kung 1990] J. P. S. Kung, "Canonical forms of binary forms: variations on a theme of Sylvester", pp. 46–58 in *Invariant theory and tableaux* (Minneapolis, MN, 1988), edited by D. Stanton, IMA Vol. Math. Appl. **19**, Springer, New York, 1990. MR 91b:11046 Zbl 0708.15013

[Kung and Rota 1984] J. P. S. Kung and G.-C. Rota, "The invariant theory of binary forms", *Bull. Amer. Math. Soc.* (*N.S.*) **10**:1 (1984), 27–85. MR 85g:05002 Zbl 0577.15020

[Landsberg 2012] J. M. Landsberg, *Tensors: geometry and applications*, Grad. Stud. Math. **128**, Amer. Math. Soc., Providence, RI, 2012. MR 2865915 Zbl 1238.15013

[Landsberg and Teitler 2010] J. M. Landsberg and Z. Teitler, "On the ranks and border ranks of symmetric tensors", *Found. Comput. Math.* **10**:3 (2010), 339–366. MR 2011d:14095 Zbl 1196.15024

[Lasker 1904] E. Lasker, "Zur Theorie der kanonischen Formen", *Math. Ann.* **58**:3 (1904), 434–440. MR 1511244 JFM 35.0122.01

[OEIS 2013] OEIS Foundation, "Online encyclopedia of integer sequences", 2013, Available at http://oeis.org.

[Pfister and Scheiderer 2012] A. Pfister and C. Scheiderer, "An elementary proof of Hilbert's theorem on ternary quartics", *J. Algebra* **371** (2012), 1–25. MR 2975385 Zbl 06166647

[Plaumann et al. 2011] D. Plaumann, B. Sturmfels, and C. Vinzant, "Quartic curves and their bitangents", *J. Symbolic Comput.* **46**:6 (2011), 712–733. MR 2012e:14065 Zbl 1214.14049

[Powers and Reznick 2000] V. Powers and B. Reznick, "Notes towards a constructive proof of Hilbert's theorem on ternary quartics", pp. 209–227 in *Quadratic forms and their applications* (Dublin, 1999), edited by E. Bayer-Fluckiger et al., Contemp. Math. **272**, Amer. Math. Soc., Providence, RI, 2000. MR 2001h:11049 Zbl 1026.11044

[Powers et al. 2004] V. Powers, B. Reznick, C. Scheiderer, and F. Sottile, "A new approach to Hilbert's theorem on ternary quartics", *C. R. Math. Acad. Sci. Paris* **339**:9 (2004), 617–620. MR 2005i:11051 Zbl 1061.11016

[Ranestad and Schreyer 2000] K. Ranestad and F.-O. Schreyer, "Varieties of sums of powers", *J. Reine Angew. Math.* **525** (2000), 147–181. MR 2001m:14009 Zbl 1078.14506

[Reichstein 1987] B. Reichstein, "On expressing a cubic form as a sum of cubes of linear forms", *Linear Algebra Appl.* **86** (1987), 91–122. MR 88e:11022 Zbl 0609.15015

[Reznick 1992a] B. Reznick, "Sums of even powers of real linear forms", *Mem. Amer. Math. Soc.* **96**:463 (1992). MR 93h:11043 Zbl 0762.11019

[Reznick 1992b] B. Reznick, "Sums of powers of complex linear forms", preprint, 1992.

[Reznick 1996] B. Reznick, "Homogeneous polynomial solutions to constant coefficient PDE's", *Adv. Math.* **117**:2 (1996), 179–192. MR 97a:12006 Zbl 0858.12004

[Reznick 2000] B. Reznick, "Some concrete aspects of Hilbert's 17th problem", pp. 251–272 in *Real algebraic geometry and ordered structures* (Baton Rouge, LA, 1996), edited by C. N. Delzell and J. J. Madden, Contemp. Math. **253**, Amer. Math. Soc., Providence, RI, 2000. MR 2001i:11042 Zbl 0972.11021

[Reznick 2013] B. Reznick, "On the length of binary forms", in *Quadratic and higher degree forms*, edited by K. Alladi et al., Dev. Math. **31**, Springer, New York, 2013. arXiv 1007.5485

[Reznick $\geq$ 2013] B. Reznick, "Linear dependence among powers of quadratic forms". In preparation.

[Richmond 1902] H. W. Richmond, "On canonical forms", *Quart. J. Pure Appl. Math.* **33** (1902), 331–340. JFM 33.0127.02

[Sylvester 1851a] J. J. Sylvester, *An essay on canonical forms: supplement to a sketch of a memoir on elimination, transformation and canonical forms*, George Bell, London, 1851. Reprinted as Paper 34, pp. 203–216 in *The collected mathematical papers of James Joseph Sylvester, I: 1837–1853*, Cambridge University Press, 2012.

[Sylvester 1851b] J. J. Sylvester, "On a remarkable discovery in the theory of canonical forms and of hyperdeterminants", *Philos. Mag.* (4) **2**:12 (1851), 391–410. Reprinted as Paper 41, pp. 265–283 in *The collected mathematical papers of James Joseph Sylvester, I: 1837–1853*, Cambridge University Press, 2012.

[Sylvester 1870] J. J. Sylvester, "Additional notes to Prof. Sylvester's Exeter British Association address", pp. 140–143 in *The laws of verse*, Longmans, London, 1870. Reprinted in Appendix, pp.

714–719 in *The collected mathematical papers of James Joseph Sylvester, II: 1857–1873*, Cambridge University Press, 2012.

[Turnbull 1960] H. W. Turnbull, *The theory of determinants, matrices, and invariants*, 3rd ed., Dover, New York, 1960. MR 24 #A123 Zbl 0103.00702

[Várilly-Alvarado 2008] A. Várilly-Alvarado, "Weak approximation on del Pezzo surfaces of degree 1", *Adv. Math.* **219**:6 (2008), 2123–2145. MR 2009j:14045 Zbl 1156.14017

[Várilly-Alvarado 2011] A. Várilly-Alvarado, "Density of rational points on isotrivial rational elliptic surfaces", *Algebra Number Theory* **5**:5 (2011), 659–690. MR 2889751 Zbl 06054517

[Wakeford 1913] E. K. Wakeford, "A canonical form of the binary sextic", *Mess. Math.* **4** (1913), 25–28. JFM 44.0142.01

[Wakeford 1920] E. K. Wakeford, "On canonical forms", *Proc. London Math. Soc.* (2) **18**:1 (1920), 403–410. MR 1576066 JFM 47.0880.01

Received March 28, 2012. Revised June 17, 2013.

BRUCE REZNICK
DEPARTMENT OF MATHEMATICS AND CENTER FOR ADVANCED STUDY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
1409 W. GREEN STREET
327 ALTGELD HALL
URBANA, IL 61801-2975
UNITED STATES

reznick@math.uiuc.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 266, No. 1, 2013
dx.doi.org/10.2140/pjm.2013.266.221

APPLICATIONS OF THE DEFORMATION FORMULA OF HOLOMORPHIC ONE-FORMS

QUANTING ZHAO AND SHENG RAO

This paper studies some geometric aspects of moduli of curves $\mathcal{M}_g$, using as a tool the deformation formula of holomorphic one-forms. Quasi-isometry guarantees the L^2 convergence of deformation of holomorphic one-forms, which is a kind of global result. After giving the period map a full expansion, we can also write out the Siegel metric, curvature and second fundamental form of a nonhyperelliptic locus of $\mathcal{M}_g$ in a quite detailed manner, while gaining some understanding of a totally geodesic manifold in a nonhyperelliptic locus.

1. Introduction

This paper is a complement to our joint paper [Liu et al. 2012b] with Kefeng Liu, and explores more applications of the deformation formula of holomorphic one-forms to some problems related to moduli spaces of Riemann surfaces, including the full expansion of the period map, the Siegel metric and its curvature formulae, the second fundamental form of Torelli space's nonhyperelliptic locus, and also a global result about the deformation of holomorphic one-forms.

We start with the Kuranishi coordinate of the Teichmüller space $\mathcal{T}_g$ of Riemann surfaces of genus g and the deformation formula of holomorphic one-forms $\theta(t)$, whose construction is contained in Section 2. The key points of the deformation formula lie in Theorem 2.1. To be more precise, on the Kuranishi family $\varpi : \mathcal{X} \to \Delta$ with a Riemann surface $\varpi^{-1}(0) = X_0$ as its central fiber and a global holomorphic one-form of the central fiber $\theta \in H^0(X_0, \Omega^1_{X_0})$, the deformation formula of holomorphic one-forms emerges as

$$\theta(t) = \theta + \sum_{|I|\geq 1} t^I \left(\sum_{j=1}^{n} \mathbb{H}\big(\mu_j \lrcorner \eta_{(i_1,\dots,i_j-1,\dots,i_n)}\big) \right), \tag{1-1}$$

Rao is the corresponding author and is partially supported by the 52nd China Postdoctoral Science Foundation, No. 2012M521157, and the National Natural Science Foundation of China, No. 11301477.
MSC2010: primary 32G15; secondary 32G20, 30F30.
Keywords: moduli of Riemann surfaces, Teichmüller theory, period matrices, variation of Hodge structures, differentials on Riemann surfaces.

such that $\theta(t) \in H^0(X_t, \Omega^1_{X_t})$, where

$$\mu(t) = \sum_{i=1}^{n} t_i \mu_i$$

is the integrable Beltrami differential of the Kuranishi family and $\mathbb{H}$ is the harmonic projector in Hodge decomposition with respect to the Poincaré metric on X_0. Moreover, $\mathbb{G}$ denotes the associated Green operator, and η_I is given by

$$\begin{cases} \eta_i = -\mathbb{G}\bar{\partial}^*\partial(\mu_i \lrcorner \theta), \\ \eta_{(i_1,\dots,i_n)} = -\mathbb{G}\bar{\partial}^*\partial\big(\sum_{k=1}^{n} \mu_k \lrcorner \eta_{(i_1,\dots,i_k-1,\dots,i_n)}\big). \end{cases}$$

We identify η_i with $\eta_{(0,\dots,\underline{1}_{i\text{-th}},\dots,0)}$ here. Apply (1-1) to the canonical basis $\{\theta_p^\alpha\}_{\alpha=1}^g$ of $H^0(X_p, \Omega^1_{X_p})$ with respect to the symplectic basis

$$\{A_\gamma, B_\gamma\}_{\gamma=1}^g$$

for the Kuranishi coordinate $\Delta_{p,\epsilon}$, yielding

$$\theta_p^\alpha(t) = \theta_p^\alpha + \sum_{i=1}^{n} t_i \big(\mathbb{H}(\mu_i \lrcorner \theta_p^\alpha) + df_i^\alpha\big) \\ + \sum_{|I|\geq 2} t^I \left(\sum_{j=1}^{n} \mathbb{H}\big(\mu_j \lrcorner \eta^\alpha_{(i_1,\dots,i_j-1,\dots,i_n)}\big) + df^\alpha_{j,(i_1,\dots,i_j-1,\dots,i_n)} \right).$$

We then define $A(t)$ by

$$\sum_{|I|\geq 1} t^I \left(\sum_{j=1}^{n} \mathbb{H}(\mu_j \lrcorner \eta^\alpha_{(i_1,\dots,i_j-1,\dots,i_n)}) \right) = A(t)^\alpha_\beta \bar{\theta}^\beta_p.$$

Let σ_p, π_p be the A, B period matrices of $\{\theta_p^\alpha\}_{\alpha=1}^g$ and $M_p = \operatorname{Im}(\pi_p)$. The period map

$$\Pi : \mathcal{T}_g \to \mathcal{H}_g,$$

where $\mathcal{H}_g$ is the classifying space of Hodge structures of weight one, can be written out on the Kuranishi coordinate as

$$\Pi(t) = (\pi_p + \bar{\pi}_p A(t)^T)(\mathbb{1}_g + A(t)^T)^{-1}. \tag{1-2}$$

H. Rauch [1959] and A. Mayer [1969] have expanded the period map only up to the first order, while Fangliang Yin's expansion formula [2010] via computing high derivatives of the period map is not explicit for orders larger than two, since it is difficult to write all the derivatives out. In a different manner, by solving a recursive relation, we can get an explicit formula for every order part of the expansion:

Theorem 1.1 (Theorem 2.5). *The period map*

$$\Pi : \mathcal{T}_g \to \mathcal{H}_g$$

has the following full expansion on the Kuranishi coordinate $\Delta_{p,\epsilon}$:

$$\begin{aligned}\Pi_{\alpha\beta}(t) = \Pi_{\alpha\beta}(0) &+ \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \theta^\beta\big) + \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^\beta_{t,1}\big) \\ &- \frac{\mathrm{i}}{2} \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \theta^\delta\big) M^{\delta\gamma} \int_{X_0} \theta^\gamma \wedge \mathbb{H}\big(\mu(t) \lrcorner \theta^\beta\big) \\ &+ \sum_{k \geq 3} \sum_{\substack{m_i > 0, 1 \leq i \leq l \\ m_1 + \cdots + m_l = k}} \Big\{ (-1)^{l-1} \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^{\alpha_1}_{t,m_1-1}\big) \\ &\qquad \times \frac{\mathrm{i}}{2} M^{\alpha_1\alpha_2} \int_{X_0} \theta^{\alpha_2} \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^{\alpha_3}_{t,m_2-1}\big) \cdots \\ &\qquad \times \frac{\mathrm{i}}{2} M^{\alpha_{2l-3}\alpha_{2l-2}} \int_{X_0} \theta^{\alpha_{2l-2}} \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^\beta_{t,m_l-1}\big) \Big\},\end{aligned}$$

where $\eta^\alpha_{t,N}$ *is the* N*-th order part of the expansion of* η^α_t *in the deformation formula of* θ^α *by Theorem 2.1,* $M^{\alpha\beta}$ *is the inverse matrix of* $M_p = \mathrm{Im}(\pi_p)$, *and* $\mathrm{i} = \sqrt{-1}$.

Geometric information on the period map is contained in the homogeneous parts in this theorem, whose meaning will become apparent in the following sections.

In Section 3, by a quasi-isometry result in [Liu et al. 2012a] for the operator $\bar\partial \circ \mathbb{G} \circ \partial$, we obtain a global result for the deformation of holomorphic one-forms.

Proposition 1.2. *The* $(1,0)$*-form* $\eta(t)$ *on* X_p *constructed in Theorem 2.1 converges in* L^2*-norm as long as* $|t| < 1$, *and so does* $\theta(t)$ *constructed in Theorem 2.1.*

In Section 4, the deformation formula of holomorphic one-forms provides us with an effective way to write out the Siegel metric and its curvature explicitly according to the expansion degree of t. To maintain that the formula is integral and clean, we need Definition 4.1 of symmetric derivatives. Also the normal coordinate (4-5) is used in our calculation.

Theorem 1.3. *The Siegel metric* $\omega_s(t)$ *on the nonhyperelliptic locus of the Torelli space* $\mathcal{T}or_g$ *can be written as*

$$\begin{aligned}\omega_s(t) &= \frac{\mathrm{i}}{2} \sum_{n=1}^{\infty} \frac{1}{n} \partial\bar\partial \ \mathrm{tr}\big(A(t)\overline{A(t)}\big)^n \\ &= \frac{\mathrm{i}}{2} \sum_{k \geq 0} \sum_{\substack{m_i > 0, 1 \leq i \leq 2l \\ m_1 + \cdots + m_{2l} = k+2}} \frac{1}{l} \mathrm{tr}\Big(\boldsymbol{S}_{i\bar j}\big(A_{m_1}(t)\overline{A_{m_2}(t)} \ldots \overline{A_{m_{2l}}(t)}\big)\Big) dt_i \wedge d\bar t_j,\end{aligned}$$

and its curvature $\mathrm{R}_{i\bar{j}k\bar{l}}$ *is given by*

$$\begin{aligned}
\mathrm{R}_{i\bar{j}k\bar{l}} = &-\sum_{N\geq 0}\sum_{\substack{m_i>0,1\leq i\leq 2l\\ m_1+\cdots+m_{2l}=N+4}} \frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\ldots\overline{A_{m_{2l}}(t)}\big)\big)\\
&+\sum_{N\geq 0}\sum_{\substack{N_i\geq 0,1\leq i\leq 3\\ \sum_{i=1}^3 N_i=N}}\Bigg[\sum_{\substack{m_i>0,1\leq i\leq l\\ m_1+\cdots+m_l=N_1}}(-1)^l\prod_{i=1}^{l}\frac{1}{s_i}\\
&\qquad\times\sum_{\substack{m_{1n}>0,1\leq n\leq 2s_1\\ \sum_{n=1}^{2s_1} m_{1n}=m_1+2}}\operatorname{tr}\big(\boldsymbol{S}_{q\bar{i_1}}\big(A_{m_{11}}(t)\ldots\overline{A_{m_{12s_1}}(t)}\big)\big)\\
&\qquad\times\sum_{\substack{m_{2n}>0,1\leq n\leq 2s_2\\ \sum_{n=1}^{2s_2} m_{2n}=m_2+2}}\operatorname{tr}\big(\boldsymbol{S}_{i_1\bar{i_2}}\big(A_{m_{21}}(t)\ldots\overline{A_{m_{22s_2}}(t)}\big)\big)\cdots\\
&\qquad\times\sum_{\substack{m_{ln}>0,1\leq n\leq 2s_l\\ \sum_{n=1}^{2s_l} m_{ln}=m_l+2}}\operatorname{tr}\big(\boldsymbol{S}_{i_{l-1}\bar{p}}\big(A_{m_{l1}}(t)\ldots\overline{A_{m_{l2s_l}}(t)}\big)\big)\Bigg]\\
&\qquad\times\Bigg[\sum_{\substack{m_i>0,1\leq i\leq 2l\\ \sum_{i=1}^{2l} m_i=N_2+3}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_i\boldsymbol{S}_{k\bar{q}}\big(A_{m_1}(t)\ldots\overline{A_{m_{2l}}(t)}\big)\big)\Bigg]\\
&\qquad\times\Bigg[\sum_{\substack{m_i>0,1\leq i\leq 2l\\ \sum_{i=1}^{2l} m_i=N_3+3}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{\bar{j}}\boldsymbol{S}_{p\bar{l}}\big(A_{m_1}(t)\ldots\overline{A_{m_{2l}}(t)}\big)\big)\Bigg],
\end{aligned}$$

where we need the convention that the first square bracket in the second summand equals δ_{qp} *if* $N_1 = 0$.

The bound $\mathrm{H}(v) \leq -2/g$ for the holomorphic sectional curvature $\mathrm{H}(v)$ along the direction $v = \sum_{i=1}^{3g-3} a_i\mu_i \in \mathbb{H}^{0,1}_{\bar\partial}(X_p, T_{X_p})$ of $\mathrm{R}_{i\bar{j}k\bar{l}}$ is also discussed in this section.

Section 5 is motivated by Oort's conjecture and Moonen's result as follows.

Conjecture 1.4 [Oort 1997]. *Let* $\overline{\mathcal{J}_g} := \overline{\mathcal{J}(\mathcal{M}_g)} \subset \mathcal{A}_g$ *be the Zariski closure of the* (*open*) *Torelli locus* $\mathcal{J}_g := \mathcal{J}(\mathcal{M}_g)$. *For* $g \geq 4$, *determine all special subvarieties* (*or varieties of Hodge type*) *of positive dimension in* $\mathcal{A}_g$ *that are contained in* $\overline{\mathcal{J}_g}$ *and meet* $\mathcal{J}_g$. *Conjecturally*, *there are no such subvarieties when g is sufficiently large.*

As a complex orbifold, $\mathcal{A}_g(\mathbb{C})$ is a quotient of the Siegel space, which is an irreducible homogeneous symmetric space under the group $\mathrm{Sp}(g, \mathbb{R})$, and special subvarieties can be considered as images of orbits of an algebraic subgroup. Here we refer the readers to the remarkable survey [Moonen and Oort 2013, Section 3.6] for the three equivalent definitions of special subvarieties and other preliminaries.

Fortunately, we have an important result by B. Moonen [1995]: Let $V \subset \mathscr{A}_g$ be an algebraic subvariety. Then V is a special subvariety if and only if it is totally geodesic with respect to the Siegel metric and it contains at least one special point. From [Oort 2003], we know that a special point in $\mathscr{A}_g$ corresponds to a moduli point of principally polarized abelian variety (A, λ) with A admitting sufficiently many complex multiplications. The notion of sufficiently many complex multiplications of an abelian variety A has an equivalent expression: there is a commutative semisimple subalgebra $E \subset \operatorname{End}^0(A) := \operatorname{End}(A) \otimes_{\mathbb{Z}} \mathbb{Q}$ with $\dim_{\mathbb{Q}} E = 2g$. Perhaps we need a more geometric description of special points for deeper investigation.

Therefore, it is important to understand the second fundamental form of the Torelli locus and totally geodesic subvarieties, which is also proposed by B. Farb [2006], R. Hain [1999] and D. Toledo [1987]. By our deformation method, we can get the second fundamental form of a nonhyperelliptic locus and some understanding of a totally geodesic manifold in a nonhyperelliptic locus. The full formula for the second fundamental form is included in the Appendix, since it is long and complicated.

Proposition 1.5. *The second fundamental form of a nonhyperelliptic locus at the central point is*

$$\Sigma_{i\bar{j}k\bar{l}}(0) = \operatorname{tr}(A_{ik}\overline{A_{jl}}),$$

where A_{ij} is defined in our discussion of the homogeneous part of $A(t)$ after Definition 2.2.

As a direct corollary, the holomorphic sectional curvature along a totally geodesic submanifold in a nonhyperelliptic locus of $\mathscr{T}_g$ is bounded from below. Moreover, we obtain the relationship between the total geodesicness and local symmetry, that is, a totally geodesic manifold in the nonhyperelliptic locus of $\mathscr{T}_g$ must be locally symmetric.

2. Full expansion of the period map

Kuranishi coordinates and small deformation of holomorphic one-forms. Fix a compact topological surface Σ of genus g with $g \geq 2$. The pair $(C, [f])$ is a Riemann surface C with the Teichmüller structure $[f]$, where f is an orientation-preserving homeomorphism from C to Σ and $[f]$ denotes the isotopic class represented by f. An isomorphism between two Riemann surfaces with the Teichmüller structures $(C, [f])$ and $(C', [f'])$ is a biholomorphic map ϕ from C to C' such that $[f] = [f'\phi]$. The equivalence classes of all compact Riemann surfaces of genus g with this Teichmüller structure, modulo the isomorphism equivalences, actually constitute $\mathscr{T}_g$. Thus an isomorphism class of $[C, [f]]$ is a point in $\mathscr{T}_g$.

From the construction of the Hilbert scheme, the existence of the Kuranishi family of Riemann surfaces follows. To be more precise, for every Riemann surface C, there exists a holomorphic deformation (ϖ, φ)

$$\varpi : \mathscr{X} \to B, \quad \varphi : C \xrightarrow{\cong} X_{b_0}$$

of C parametrized by a pointed base (B, b_0) and a complex manifold with $\dim_{\mathbb{C}} B = 3g-3$; this deformation is universal at b_0 and actually universal at every point b of B. The pair (ϖ, φ) is called the Kuranishi family of C. For any other deformation (ι, ψ)

$$\iota : \mathscr{X}' \to B', \quad \psi : C \xrightarrow{\cong} X'_{b'_0}$$

of C, there exists a unique map (ϕ, Φ) in a small neighborhood of b'_0 such that the diagram

$$\begin{tikzcd}
\mathscr{X}' \arrow[r, "\Phi"] \arrow[d, "\iota"'] & \mathscr{X} \arrow[d, "\varpi"] \\
(B', b'_0) \arrow[r, "\phi"] & (B, b_0)
\end{tikzcd}$$

commutes, where $\varphi^{-1}\Phi_{b'_0}\psi = \mathbb{1}_C$ and $\mathscr{X}'$ is isomorphic to the pullback family $\Phi^*\mathscr{X}$ on the small neighborhood of b'_0. Accordingly, we also have a family of Riemann surfaces with the Teichmüller structure $(X_b, [f_b])$, that is, $\varpi : \mathscr{X} \to B$, together with the local topological trivialization

$$F^\alpha : \mathscr{X}|_{U_\alpha} \to \Sigma \times U_\alpha,$$

where $\bigcup_\alpha U_\alpha$ is an open covering of B such that $[F_b^\alpha] = [f_b]$ with $b \in U_\alpha$. For any Riemann surface with the Teichmüller structure $(C, [f])$, the Kuranishi family also exists and satisfies exactly analogous universal properties to the one without this Teichmüller structure. Possibly after shrinking B, we can describe the Kuranishi family of $(C, [f])$ as a triple (ϖ, φ, F) given by

$$\varpi : \mathscr{X} \to B, \quad \varphi : C \xrightarrow{\cong} C_{b_0}, \quad F : \mathscr{X} \to \Sigma \times B,$$

where F is a topological trivialization such that $F_{b_0}\varphi = f$.

A Kuranishi coordinate chart of $\mathscr{T}_g$ is given by

$$(B, b_0) \to \mathscr{T}_g, \quad t \to [X_t, [F_t]],$$

where the triple (ϖ, φ, F) is the Kuranishi family of $(C, [f])$. By Ehresmann's classical theorem, there is a natural diffeomorphism $\Psi : X_{b_0} \times B \to \mathscr{X}$; all the fibers of $\varpi : \mathscr{X} \to B$

$$\begin{tikzcd}
\Sigma \times B & \\
\mathscr{X} \arrow[u, "F"] \arrow[r, "\Psi"] & X_{b_0} \times B
\end{tikzcd}$$

share the same differential structure as X_{b_0}. From this point of view, for every $b \in B$, the map $F_b \Psi_b^{-1}$ can be deformed to $F_{b_0} \Psi_{b_0}^{-1}$, that is,

$$[F_b \Psi_b^{-1}] = [F_{b_0} \Psi_{b_0}^{-1}].$$

Let $\omega\colon H_1(\Sigma, \mathbb{Z}) \times H_1(\Sigma, \mathbb{Z}) \to \mathbb{Z}$ be the intersection pairing on Σ. The symplectic basis of $H_1(\Sigma, \mathbb{Z})$ on (Σ, ω) gives, from the map ΨF^{-1}, one such basis on X_{b_0}, which is enjoyed by the whole Kuranishi family $\mathscr{X}$ over the Kuranishi coordinate chart B. Later on we will write (B, b_0) as $\Delta_{p,\epsilon}$, where p denotes the point $[C, [f]]$ in $\mathscr{T}_g$, and $\Delta_{p,\epsilon} = \left\{t \in \mathbb{C}^n \,\middle|\, \|t\| < \epsilon,\ t(p) = 0\right\}$ with $n = 3g - 3$.

Let $\Delta_{p,\epsilon}$ be the Kuranishi coordinate centered at $p \in \mathscr{T}_g$ above. Denote the corresponding Kuranishi family on $\Delta_{p,\epsilon}$ by $\varpi : \mathscr{X} \to \Delta_{p,\epsilon}$ with the central fiber $\varpi^{-1}(0) = X_p$. Let

$$\theta \in H^0(X_p, \Omega^1_{X_p})$$

be a global holomorphic one-form on X_p. We will construct $\theta(t) \in H^0(X_t, \Omega^1_{X_t})$, a holomorphic deformation of θ with t small.

Denote the well known Poincaré metric on X_p by ω_p. Fix $\{\mu_i\}_{i=1}^n$ as a basis of harmonic $T^{(1,0)}_{X_p}$-valued (0,1) forms, written as $\mathbb{H}^{0,1}_{\bar\partial}(X_p, T^{(1,0)}_{X_p})$, on (X_p, ω_p). And $\mu(t) = \sum_{i=1}^n t_i \mu_i$ is the integrable Beltrami differential of the Kuranishi family $\varpi : \mathscr{X} \to \Delta_{p,\epsilon}$.

Theorem 2.1 [Liu et al. 2012b, Theorem 2.1 and Corollary 2.2]. *Given $\theta \in H^0(X_p, \Omega^1_{X_p})$, there exists a unique $(1,0)$-form $\eta(t)$ on X_p that is holomorphic in t for sufficiently small t, satisfying*

(1) *$\mathbb{H}(\eta(t)) = \theta$, where $\mathbb{H}$ is the harmonic projection on (X_p, ω_p), and*

(2) *$\theta(t) = (\mathbb{1} + \mu(t)) \lrcorner \eta(t) \in H^0(X_t, \Omega^1_{X_t})$,*

and $\theta(t)$ is the desired deformation of θ, given by

$$\theta(t) = \theta + \sum_{i=1}^n t_i \left(\mathbb{H}(\mu_i \lrcorner \theta) + df_i\right) \\ + \sum_{|I| \ge 2} t^I \left(\sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta_{(i_1,\dots,i_j-1,\dots,i_n)}\big) + df_{j,(i_1,\dots,i_j-1,\dots,i_n)} \right),$$

where $f_{j,(i_1,\dots,i_j-1,\dots,i_n)} \in C^\infty(X_p)$.

Based on this, an explicit formula of the period map and variation of Hodge structures on Kuranishi coordinates are discussed below. Denote the canonical basis of $H^0(X_p, \Omega^1_{X_p})$ by $\{\theta_p^\alpha\}_{\alpha=1}^g$ with respect to the symplectic basis $\{A_\gamma, B_\gamma\}_{\gamma=1}^g$ for the Kuranishi coordinate $\Delta_{p,\epsilon}$. Let σ_p, π_p be the A, B period matrices of

$\{\theta_p^\alpha\}_{\alpha=1}^g$ and $M_p = \mathrm{Im}(\pi_p)$. Applying the deformation formula above, we get the holomorphic one-forms $\theta_p^\alpha(t)$ on X_t, starting at θ_p^α, given by

$$\begin{aligned}\theta_p^\alpha(t) = \theta_p^\alpha + \sum_{i=1}^n t_i\big(\mathbb{H}(\mu_i \lrcorner \theta_p^\alpha) + df_i^\alpha\big) \\ + \sum_{|I|\geq 2} t^I\Bigg(\sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta^\alpha_{(i_1,\dots,i_j-1,\dots,i_n)}\big) + df^\alpha_{j,(i_1,\dots,i_j-1,\dots,i_n)}\Bigg).\end{aligned} \tag{2-1}$$

Definition 2.2 ($A(t)$ and $E(t)$). $A(t)$ is a $g\times g$ matrix, while $E(t)$ is a $g\times 1$ vector defined by

$$\sum_{|I|\geq 1} t^I \sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta^\alpha_{(i_1,\dots,i_j-1,\dots,i_n)}\big) = A(t)^\alpha_\beta \bar{\theta}^\beta_p,$$

$$\sum_{|I|\geq 1} t^I \Bigg(\sum_{j=1}^n df^\alpha_{j,(i_1,\dots,i_j-1,\dots,i_n)}\Bigg) = E^\alpha(t).$$

We write the homogeneous part of order N of $A(t)$ as $A_N(t)$. Then

$$A_N(t) = \sum_{|I|=N} t^I A_I,$$

where $\sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta^\alpha_{(i_1,\dots,i_j-1,\dots,i_n)}\big) = A_I{}_\beta^\alpha\, \bar{\theta}^\beta_p$. In particular, $\mathbb{H}(\mu_i \lrcorner \theta_p^\alpha) = A_i{}_\beta^\alpha\, \bar{\theta}^\beta_p$.

Detailed discussion of homogeneous parts of $A(t)$ is given as follows, and will be useful in the computation.

(1) The first two homogeneous parts $A_1(t)$ and $A_2(t)$:

$$\mathbb{H}(\mu_i \lrcorner \theta_p^\alpha) = A_i{}_\beta^\alpha\, \bar{\theta}^\beta_p \quad \text{for } |I| = 1,$$

$$\begin{cases}\mathbb{H}(\mu_i \lrcorner \eta_j^\alpha) + \mathbb{H}(\mu_j \lrcorner \eta_i^\alpha) = A_{(0,\dots,0,\underline{1}_{i\text{-th}},0,\dots,0,\underline{1}_{jth},0,\dots,0)}{}_\beta^\alpha\, \bar{\theta}^\beta_p & \text{for } |I|=2,\, i\neq j,\\ \mathbb{H}(\mu_i \lrcorner \eta_i^\alpha) = A_{(0,\dots,0,\underline{2}_{i\text{-th}},0,\dots,0)}{}_\beta^\alpha\, \bar{\theta}^\beta_p & \text{for } |I|=2,\, i=j.\end{cases}$$

Set

$$A_{ij} := \begin{cases} A_{(0,\dots,0,\underline{1}_{i\text{-th}},0,\dots,0,\underline{1}_{jth},0,\dots,0)} & \text{for } i<j,\\ 2A_{(0,\dots,0,\underline{2}_{i\text{-th}},0,\dots,0)} & \text{for } i=j,\\ A_{ji} & \text{for } i>j.\end{cases}$$

Then it is easy to check that

$$\sum_{i,j=1}^n t_i t_j A_{ij} = 2\sum_{|I|=2} t^I A_I \quad \text{and} \quad \frac{\partial}{\partial t_i}\frac{\partial}{\partial t_j} A_2(t) = A_{ij}.$$

(2) The general homogeneous part $A_N(t)$:

Define $A_{i_1,\dots,i_N}$, symmetric in all its subscripts, such that

$$\sum_{i_1,\dots,i_N=1}^{n} t_{i_1}\dots t_{i_N} A_{i_1,\dots,i_N} = N!\,A_N(t).$$

From this, we can deduce that

$$\frac{N!}{j_1!\,j_2!\dots j_n!} A_{i_1,\dots,i_N} = N!\,A_{(j_1,j_2,\dots,j_n)}$$

if k appears j_k times in $i_1,\dots,i_N$. Here $\sum_{k=1}^{n} j_k = N$ since we are considering the homogeneous N-part. Also it is easy to check that

$$\frac{\partial}{\partial t_{i_1}}\dots\frac{\partial}{\partial t_{i_N}} A_N(t) = A_{i_1,\dots,i_N}.$$

Set

$$\Theta_p(t) = \begin{pmatrix} \theta_p^1(t) \\ \vdots \\ \theta_p^g(t) \end{pmatrix} \quad\text{and}\quad \Theta_p = \begin{pmatrix} \theta_p^1 \\ \vdots \\ \theta_p^g \end{pmatrix}.$$

Thus using $A(t)$ and $E(t)$, we rewrite (2-1) as

$$\Theta_p(t) = (\mathbb{1}_g \;\; A(t))\begin{pmatrix} \Theta_p \\ \bar{\Theta}_p \end{pmatrix} + E(t). \tag{2-2}$$

Since a holomorphic one-form on a Riemann surface is uniquely determined by its integration on A cycles, it is clear that $\{\theta_p^\alpha(t)\}_{\alpha=1}^g$ being a frame of $H^0(X_t,\Omega^1_{X_t})$ on X_t is equivalent to nondegeneracy of the A period matrix $\sigma_{\alpha\beta}(t)$ on X_t, i.e.,

$$\det(\sigma_{\alpha\beta}(t)) = \det\left(\int_{A_\alpha} \theta_p^\beta(t)\right) \neq 0 \iff \det(\mathbb{1}_g + A(t)^T) \neq 0, \tag{2-3}$$

where $A(t)^T$ is the transpose of $A(t)$. And when $\{\theta_p^\alpha(t)\}_{\alpha=1}^g$ becomes a frame, we have the Hodge–Riemann bilinear relations on X_t

$$\begin{cases} 0 = \dfrac{\mathrm{i}}{2}\displaystyle\int_{X_t} \theta_p^\alpha(t)\wedge\theta_p^\beta(t), \\ 0 < \dfrac{\mathrm{i}}{2}\displaystyle\int_{X_t} \theta_p^\alpha(t)\wedge\bar{\theta}_p^\beta(t), \end{cases}$$

which, together with (2-2), implies that

$$\begin{cases} 0 = \dfrac{\mathrm{i}}{2}\displaystyle\int_{X_p} \big(\theta_p^\alpha + A(t)^\alpha_\gamma\bar{\theta}^\gamma + E^\alpha(t)\big)\wedge\big(\theta_p^\beta + A(t)^\beta_\lambda\bar{\theta}^\lambda + E^\beta(t)\big), \\ 0 < \dfrac{\mathrm{i}}{2}\displaystyle\int_{X_p} \big(\theta_p^\alpha + A(t)^\alpha_\gamma\bar{\theta}_p^\gamma + E^\alpha(t)\big)\wedge\big(\bar{\theta}_p^\beta + \overline{A(t)}^\beta_\lambda\theta_p^\lambda + \bar{E}^\beta(t)\big). \end{cases}$$

Thus, by type consideration and Stokes's theorem, we have

$$\begin{cases} 0 = M_{p,\alpha\gamma}\, A(t)^{\beta}_{\gamma} - M_{p,\beta\gamma}\, A(t)^{\alpha}_{\gamma}, \\ 0 < M_{p,\alpha\beta} - M_{p,\lambda\gamma}\, A(t)^{\alpha}_{\gamma} \overline{A(t)}^{\beta}_{\lambda}. \end{cases}$$

The matrix forms of these are given by

$$\begin{cases} A(t)M_p = (A(t)M_p)^T, \\ M_p - A(t)M_p\overline{A(t)}^T > 0. \end{cases} \tag{2-4}$$

As our deformation formula is local, $\{\theta^{\alpha}_p(t)\}_{\alpha=1}^{g}$ is always a frame, as $t \in \Delta_{p,\epsilon}$ with ϵ sufficiently small. Therefore, (2-3) and (2-4) hold.

On our Kuranishi coordinate $\Delta_{p,\epsilon}$, the period map $\Pi : \mathcal{T}_g \to \mathcal{H}_g$ can be written out quite explicitly:

$$\begin{aligned} \Pi(t)_{\alpha\beta} &= \int_{B_\alpha} \sigma(t)^{\gamma\beta}\theta^{\gamma}_p(t) = \int_{B_\alpha} \sigma(t)^{\gamma\beta}\big(\theta^{\gamma}_p + A(t),^{\gamma}_{\delta}\, \bar{\theta}^{\delta}_p\big) \\ &= \pi_{p,\alpha\gamma}\, \sigma(t)^{\gamma\beta} + \bar{\pi}_{p,\alpha\delta}\, A(t)^{\gamma}_{\delta}\sigma(t)^{\gamma\beta}, \end{aligned} \tag{2-5}$$

where $\sigma(t)^{\alpha\beta}$ is the inverse matrix of $\sigma(t)_{\alpha\beta}$. Here $\sigma(t)_{\alpha\beta}$ is given by

$$\sigma_{\alpha\beta}(t) = \int_{A_\alpha} \theta^{\beta}_p(t) = (\mathbb{1}_g + A(t)^T)_{\alpha\beta}. \tag{2-6}$$

By (2-6), we formulate (2-5) into the matrix type to get

$$\Pi(t) = (\pi_p + \bar{\pi}_p A(t)^T)(\mathbb{1}_g + A(t)^T)^{-1}. \tag{2-7}$$

Full expansion of the period map. We are going to give (2-7) a full expansion, writing out every order part explicitly.

Lemma 2.3 [Farkas and Kra 1992, Proposition III.2.3]. *If ϕ and ψ are two d-closed one-forms on a Riemann surface X, then*

$$\int_X \phi \wedge \psi = \sum_{\gamma} \Big(\int_{A_\gamma} \phi \int_{B_\gamma} \psi - \int_{B_\gamma} \phi \int_{A_\gamma} \psi \Big),$$

where $\{A_\gamma, B_\gamma\}_{\gamma=1}^{g}$ is the symplectic basis of X.

Lemma 2.4. *We hve*

$$\int_{A_\alpha} \mathbb{H}(\mu_k \lrcorner \theta^{\beta}) = \frac{\mathrm{i}}{2} M^{\alpha\gamma} \int_X \theta^{\gamma} \wedge \mathbb{H}(\mu_k \lrcorner \theta^{\beta}),$$

where $\{\theta^{\alpha}\}_{\alpha=1}^{g}$ is the canonical basis of holomorphic one-forms on X and $M^{\alpha\beta}$ is the inverse matrix of $M_{\alpha\beta} = \mathrm{Im}(\pi_{\alpha\beta})$.

Proof. Set $\mathbb{H}(\mu_k \lrcorner \theta^\beta) = c^\beta_{k,\gamma}\bar{\theta}^\gamma$. Then

$$\int_{A_\alpha} \mathbb{H}(\mu_k \lrcorner \theta^\beta) = c^\beta_{k,\alpha},$$

while Lemma 2.3 implies that

$$\mathrm{i}\int_{X_0} \theta^\gamma \wedge \mathbb{H}(\mu_k \lrcorner \theta^\beta) = \mathrm{i}c^\beta_{k,\gamma}\int_{X_0} \theta^\alpha \wedge \bar{\theta}^\gamma = 2c^\beta_{k,\gamma} M_{\alpha\gamma}.$$

Finally we have the equality above. □

Theorem 2.5. *The period map $\Pi : \mathcal{T}_g \to \mathcal{H}_g$ has the full expansion on the Kuranishi coordinate $\Delta_{p,\epsilon}$*

$$\begin{aligned}
\Pi_{\alpha\beta}(t) = \Pi_{\alpha\beta}(0) &+ \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \theta^\beta\big) + \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^\beta_{t,1}\big) \\
&- \frac{\mathrm{i}}{2}\int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \theta^\delta\big) M^{\delta\gamma} \int_{X_0} \theta^\gamma \wedge \mathbb{H}\big(\mu(t) \lrcorner \theta^\beta\big) \\
&+ \sum_{k\ge 3} \sum_{\substack{m_i>0,1\le i\le l \\ m_1+\cdots+m_l=k}} \Big\{(-1)^{l-1}\int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^{\alpha_1}_{t,m_1-1}\big) \\
&\qquad\qquad \times \frac{\mathrm{i}}{2} M^{\alpha_1\alpha_2}\int_{X_0} \theta^{\alpha_2} \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^{\alpha_3}_{t,m_2-1}\big) \cdots \\
&\qquad\qquad \times \frac{\mathrm{i}}{2} M^{\alpha_{2l-3}\alpha_{2l-2}}\int_{X_0} \theta^{\alpha_{2l-2}} \wedge \mathbb{H}\big(\mu(t) \lrcorner \eta^{\beta}_{t,m_l-1}\big)\Big\},
\end{aligned}$$

where $\eta^\alpha_{t,N}$ is the N-th order part of the expansion of η^α_t from the deformation formula of θ^α by Theorem 2.1 and $M^{\alpha\beta}$ is the inverse matrix of $M_p = \mathrm{Im}(\pi_p)$.

Proof. Write out A, B periods of X_t as $\sigma_{\alpha\beta}(t)$, $\pi_{\alpha\beta}(t)$, respectively. Then

$$\sigma_{\alpha\beta}(t) = \int_{A_\alpha} \theta^\beta(t) = \int_{A_\alpha}\bigg(\theta^\beta + \sum_{|I|\ge 1} t^I\bigg(\sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta^\beta_{(i_1,\dots,i_j-1,\dots,i_n)}\big)\bigg)\bigg),$$

$$\pi_{\alpha\beta}(t) = \int_{B_\alpha} \theta^\beta(t) = \int_{B_\alpha}\bigg(\theta^\beta + \sum_{|I|\ge 1} t^I\bigg(\sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta^\beta_{(i_1,\dots,i_j-1,\dots,i_n)}\big)\bigg)\bigg).$$

Those expansion coefficients are

$$\begin{cases} \sigma_{\alpha\beta,I} = \displaystyle\int_{A_\alpha} \sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta^\beta_{(i_1,\dots,i_j-1,\dots,i_n)}\big), \\ \pi_{\alpha\beta,I} = \displaystyle\int_{B_\alpha} \sum_{j=1}^n \mathbb{H}\big(\mu_j \lrcorner \eta^\beta_{(i_1,\dots,i_j-1,\dots,i_n)}\big). \end{cases} \tag{2-8}$$

Thus the period map can be computed as

$$\Pi_{\alpha\beta}(t)=\int_{B_\alpha}\tilde{\theta}_t^\beta=\int_{B_\alpha}\sigma^{\gamma\beta}(t)\theta_t^\gamma=\sigma^{\gamma\beta}(t)\pi_{\alpha\gamma}(t), \tag{2-9}$$

where $\tilde{\theta}^\beta(t)$ is the canonical basis of holomorphic one-forms on X_t and $\sigma^{\beta\alpha}(t)$ is the inverse matrix of $\sigma_{\alpha\beta}(t)$.

Now we only focus on the expansion of $\sigma^{\gamma\beta}(t)$, since the expansion of $\pi_{\alpha\beta}(t)$ is already obtained. Set

$$\sigma^{\alpha\beta}(t)=\delta_{\alpha\beta}+\sum_{|I|\geq 1}t^I\sigma_I^{\alpha\beta};$$

then

$$\begin{aligned}\delta_{\alpha\gamma}=\sigma_{\alpha\beta}(t)\sigma^{\beta\gamma}(t)&=\Big(\delta_{\alpha\beta}+\sum_{|I|\geq 1}t^I\sigma_{\alpha\beta,I}\Big)\Big(\delta_{\beta\gamma}+\sum_{|I|\geq 1}t^I\sigma_I^{\beta\gamma}\Big)\\&=\delta_{\alpha\gamma}+t^I\sigma_{\alpha\gamma,I}+t^I\sigma_I^{\alpha\gamma}+t^{I+J}\sigma_{\alpha\beta,I}b_J^{\beta\gamma}\\&=\delta_{\alpha\gamma}+\sum_{i=1}^n t_i\big(\sigma_{\alpha\gamma,(0,\dots,\underset{i\text{-th}}{1},\dots,0)}+\sigma^{\alpha\gamma}_{(0,\dots,\underset{i\text{-th}}{1},\dots,0)}\big)\\&\qquad+\sum_{|K|\geq 2}t^K\Big(\sigma_{\alpha\gamma,K}+\sigma_K^{\alpha\gamma}+\sum_{\substack{|I|\geq 1,|J|\geq 1\\ I+J=K}}\sigma_{\alpha\beta,I}\sigma_J^{\beta\gamma}\Big).\end{aligned}$$

Compare both sides of this equation to get

$$\begin{cases}\sigma_{\alpha\gamma,(0,\dots,\underline{1}_{i\text{-th}},\dots,0)}+\sigma^{\alpha\gamma}_{(0,\dots,\underset{i\text{-th}}{1},\dots,0)}=0,\\ \sigma_K^{\alpha\gamma}+\sigma_{\alpha\gamma,K}+\sum_{\substack{|I|\geq 1,|J|\geq 1\\ I+J=K}}\sigma_{\alpha\beta,I}\sigma_J^{\beta\gamma}=0.\end{cases} \tag{2-10}$$

Define the homogeneous parts of $\sigma_{\alpha\beta}(t)$, $\pi_{\alpha\beta}(t)$ and $\sigma^{\alpha\beta}(t)$ as

$$(\sigma_{\alpha\beta})_k:=\sum_{|K|=k}t^K\sigma_{\alpha\beta,K},\quad (\pi_{\alpha\beta})_k:=\sum_{|K|=k}t^K\pi_{\alpha\beta,K},\quad (\sigma^{\alpha\beta})_k:=\sum_{|K|=k}t^K\sigma_K^{\alpha\beta}. \tag{2-11}$$

Using these definitions, we rewrite (2-10) to obtain the recursive relation

$$\begin{cases}(\sigma_{\alpha\gamma})_1+(\sigma^{\alpha\gamma})_1=0,\\ (\sigma_{\alpha\gamma})_k+(\sigma^{\alpha\gamma})_k+\sum_{\substack{i\geq 1,j\geq 1\\ i+j=k}}(\sigma_{\alpha\beta})_i(\sigma^{\beta\gamma})_j=0.\end{cases} \tag{2-12}$$

From (2-9) and (2-11), we get

$$\begin{aligned}\Pi_{\alpha\beta}(t)=\sigma^{\gamma\beta}(t)\pi_{\alpha\gamma}(t)&=\Big(\delta_{\gamma\beta}+\sum_{|I|\ge 1}t^I\sigma_I^{\gamma\beta}\Big)\Big(\pi_{\alpha\gamma}(0)+\sum_{|I|\ge 1}t^I\pi_{\alpha\gamma,I}\Big)\\&=\pi_{\alpha\beta}(0)+\sum_{i=1}^n t_i\big(\pi_{\alpha\gamma}(0)\sigma^{\gamma\beta}_{(0,\dots,\underline{1}_{i\text{-th}},\dots,0)}+\pi_{\alpha\beta,(0,\dots,\underline{1}_{i\text{-th}},\dots,0)}\big)\\&\qquad+\sum_{|K|\ge 2}t^K\Big\{\pi_{\alpha\beta,K}+\pi_{\alpha\gamma}(0)\sigma_K^{\gamma\beta}+\sum_{\substack{|I|\ge 1,|J|\ge 1\\ I+J=K}}\pi_{\alpha\gamma,I}\sigma_J^{\gamma\beta}\Big\}\\&=\Pi_{\alpha\beta}(0)+\pi_{\alpha\gamma}(0)(\sigma^{\gamma\beta})_1+(\pi_{\alpha\beta})_1\\&\qquad+\sum_{k\ge 2}\Big\{(\pi_{\alpha\beta})_k+\pi_{\alpha\gamma}(0)(\sigma^{\gamma\beta})_k+\sum_{\substack{i\ge 1,j\ge 1\\ i+j=k}}(\pi_{\alpha\gamma})_i(\sigma^{\gamma\beta})_j\Big\}.\end{aligned}$$

After observing the formula above, we need to use the recursion relation (2-12) to get the full expansion of $\sigma^{\alpha\beta}(t)$.

Claim. *$\sigma^{\alpha\beta}(t)$ has the expansion*

$$\begin{aligned}\sigma^{\alpha\beta}(t)=\delta_{\alpha\beta}&-\int_{A_\alpha}\mathbb{H}(\mu(t)\lrcorner\theta^\beta)\\&+\sum_{k\ge 2}\sum_{\substack{m_i>0,1\le i\le n\\ m_1+\cdots+m_l=k}}\Big\{(-1)^l\int_{A_\alpha}\mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_1}_{t,m_1-1}\big)\int_{A_{\alpha_1}}\mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_2}_{t,m_2-1}\big)\cdots\\&\qquad\times\int_{A_{\alpha_{l-2}}}\mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_{l-1}}_{t,m_{l-1}-1}\big)\int_{A_{\alpha_{l-1}}}\mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,m_l-1}\big)\Big\}.\end{aligned}$$

Proof. Use an induction argument: For $k=1$, according to (2-12), the first-order part of $\sigma^{\alpha\beta}(t)$ is given by

$$(\sigma^{\alpha\beta})_1=-(\sigma_{\alpha\beta})_1=-\int_{A_\alpha}\mathbb{H}(\mu(t)\lrcorner\theta^\beta).$$

Assume that the homogeneous parts with orders less than or equal to $k-1$ are given by the formula in the claim. Then the k-th term is

$$\begin{aligned}(\sigma^{\alpha\beta})_k&=-(\sigma_{\alpha\beta})_k-\sum_{\substack{i\ge 1,j\ge 1\\ i+j=k}}(\sigma_{\alpha\gamma})_i(\sigma^{\gamma\beta})_j\\&=-\int_{A_\alpha}\mathbb{H}\big(\mu(t)\lrcorner\eta^\beta_{t,k-1}\big)-\sum_{\substack{i\ge 1,j\ge 1\\ i+j=k}}\int_{A_\alpha}\mathbb{H}\big(\mu(t)\lrcorner\eta^\gamma_{t,i-1}\big)\\&\qquad\times\Big\{\sum_{\substack{m_i>0,1\le i\le l\\ m_1+\cdots+m_l=j}}(-1)^l\int_{A_\gamma}\mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_1}_{t,m_1-1}\big)\cdots\int_{A_{\alpha_{l-1}}}\mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,m_l-1}\big)\Big\}\end{aligned}$$

$$
\begin{aligned}
&= -\int_{A_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,k-1}\big) + \sum_{\substack{i\neq k, m_j>0, 1\leq j\leq l\\ i+m_1+\cdots+m_l=k}} (-1)^{l+1}\int_{A_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\gamma}_{t,i-1}\big)\\
&\qquad\qquad \times \int_{A_\gamma} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_1}_{t,m_1-1}\big)\cdots\int_{A_{\alpha_{l-1}}} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,m_l-1}\big)\\
&= -\int_{A_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,k-1}\big) + \sum_{\substack{m_1\neq k, m_i>0, 2\leq i\leq l\\ m_1+\cdots+m_l=k}} (-1)^{l}\int_{A_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_1}_{t,m_1-1}\big)\\
&\qquad\qquad \times \int_{A_{\alpha_1}} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_2}_{t,m_1-1}\big)\cdots\int_{A_{\alpha_{l-1}}} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,m_l-1}\big)\\
&= \sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=k}} (-1)^{l}\int_{A_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_1}_{t,m_1-1}\big)\\
&\qquad\qquad \times \int_{A_{\alpha_1}} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_2}_{t,m_1-1}\big)\cdots\int_{A_{\alpha_{l-1}}} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,m_l-1}\big).
\end{aligned}
$$

Thus our claim has been proved. □

Let us proceed to the expansion of $\Pi_{\alpha\beta}(t)$. Use the claim and the expansion formula of $\Pi_{\alpha\beta}(t)$ above to get

$$
\begin{aligned}
&\Pi_{\alpha\beta}(t)\\
&= \Pi_{\alpha\beta}(0) + \pi_{\alpha\gamma}(0)(\sigma^{\gamma\beta})_1 + \int_{B_\alpha} \mathbb{H}(\mu(t)\lrcorner\theta^{\beta})\\
&\quad + \sum_{k\geq 2}\Big\{\int_{B_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,k-1}\big) + \pi_{\alpha\gamma}(0)(\sigma^{\gamma\beta})_k + \sum_{\substack{i\geq 1, j\geq 1\\ i+j=k}} \int_{B_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\gamma}_{t,i-1}\big)(\sigma^{\gamma\beta})_j\Big\}\\
&= \Pi_{\alpha\beta}(0) - \pi_{\alpha\gamma}(0)\int_{A_\gamma} \mathbb{H}(\mu(t)\lrcorner\theta^{\beta}) + \int_{B_\alpha} \mathbb{H}(\mu(t)\lrcorner\theta^{\beta})\\
&\quad + \sum_{k\geq 2}\Big\{\int_{B_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,k-1}\big) - \pi_{\alpha\gamma}(0)\int_{A_\gamma} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,k-1}\big)\\
&\qquad - \sum_{\substack{i\geq 1, j\geq 1\\ i+j=k}} \pi_{\alpha\gamma}(0)\int_{A_\gamma} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\sigma}_{t,i-1}\big)(\sigma^{\sigma\beta})_j + \sum_{\substack{i\geq 1, j\geq 1\\ i+j=k}} \int_{B_\alpha} \mathbb{H}\big(\mu(t)\lrcorner\eta^{\sigma}_{t,i-1}\big)(\sigma^{\sigma\beta})_j\Big\}\\
&= \Pi_{\alpha\beta}(0) + \int_{X_0} \theta^{\alpha}\wedge \mathbb{H}(\mu(t)\lrcorner\theta^{\beta})\\
&\qquad + \sum_{k\geq 2}\Big\{\int_{X_0} \theta^{\alpha}\wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,k-1}\big) + \sum_{\substack{i\geq 1, j\geq 1\\ i+j=k}} \int_{X_0} \theta^{\alpha}\wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^{\sigma}_{t,i-1}\big)(\sigma^{\sigma\beta})_j\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&= \Pi_{\alpha\beta}(0) + \int_{X_0} \theta^\alpha \wedge \mathbb{H}(\mu(t) \lrcorner \theta^\beta) \\
&\quad + \sum_{k\geq 2} \bigg\{ \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^\beta_{t,k-1}\big) + \sum_{\substack{i\geq 1, j\geq 1\\ i+j=k}} \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^\sigma_{t,i-1}\big) \\
&\qquad \times \Big[\sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=j}} (-1)^l \int_{A_\sigma} \mathbb{H}\big(\mu(t)\lrcorner \eta^{\alpha_1}_{t,m_1-1}\big) \cdots \int_{A_{\alpha_{l-1}}} \mathbb{H}\big(\mu(t)\lrcorner \eta^\beta_{t,m_l-1}\big) \Big] \bigg\} \\
&= \Pi_{\alpha\beta}(0) + \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t) \lrcorner \theta^\beta\big) \\
&\quad + \sum_{k\geq 2} \bigg\{ \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^\beta_{t,k-1}\big) + \sum_{\substack{i\neq k, m_j>0, 1\leq j\leq l\\ i+m_1+\cdots+m_l=k}} (-1)^l \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^\sigma_{t,i-1}\big) \\
&\qquad \times \int_{A_\sigma} \mathbb{H}\big(\mu(t)\lrcorner \eta^{\alpha_1}_{t,m_1-1}\big) \cdots \int_{A_{\alpha_{l-1}}} \mathbb{H}\big(\mu(t)\lrcorner \eta^\beta_{t,m_l-1}\big) \bigg\} \\
&= \Pi_{\alpha\beta}(0) + \int_{X_0} \theta^\alpha \wedge \mathbb{H}(\mu(t) \lrcorner \theta^\beta) \\
&\quad + \sum_{k\geq 2} \sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=k}} \bigg\{ (-1)^{l-1} \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^{\alpha_1}_{t,m_1-1}\big) \times \\
&\quad \frac{\mathrm{i}}{2} M^{\alpha_1\alpha_2} \int_{X_0} \theta^{\alpha_2} \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^{\alpha_3}_{t,m_2-1}\big) \cdots \frac{\mathrm{i}}{2} M^{\alpha_{2l-3}\alpha_{2l-2}} \int_{X_0} \theta^{\alpha_{2l-2}} \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^\beta_{t,m_l-1}\big) \bigg\}
\end{aligned}
$$

and this concludes the proof of Theorem 2.5. □

Corollary 2.6. *For every* $N \geq 0$,

$$
\int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner \eta^\beta_{t,N}\big)
$$

is a symmetric matrix of (α, β).

Proof. We again use an induction argument.

When $N = 0$, $\int_{X_0} \theta^\alpha \wedge \mathbb{H}(\mu(t)\lrcorner\theta^\beta)$ is the homogeneous part of first order of $\Pi_{\alpha\beta}(t)$. It is easy to see that $\Pi_{\alpha\beta}(t)$ is a symmetric matrix of (α, β), and thus the homogeneous part of every order of its expansion will be symmetric in (α, β), and in particular the first order.

Assume that $\int_{X_0} \theta^\alpha \wedge \mathbb{H}(\mu(t)\lrcorner\eta^\beta_{t,N})$, with $N \leq k-1$, is symmetric in (α, β). Now we come to the case $N = k$. By Theorem 2.5, the $(k+1)$-th homogeneous part of the expansion of $\Pi_{\alpha\beta}$ is

$$\int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^\beta_{t,k}\big) + \sum_{\substack{m_i \le k \\ m_1+m_2+\cdots+m_l=k+1}} (-1)^{l-1} \int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_1}_{t,m_1-1}\big) \times \frac{\mathrm{i}}{2} M^{\alpha_1\alpha_2} \int_{X_0} \theta^{\alpha_2} \wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^{\alpha_3}_{t,m_2-1}\big) \cdots \times \frac{\mathrm{i}}{2} M^{\alpha_{2l-3}\alpha_{2l-2}} \int_{X_0} \theta^{\alpha_{2l-2}} \wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^{\beta}_{t,m_l-1}\big)$$

and is thus symmetric in (α, β). By use of the induction assumption and the symmetric matrix $M^{\alpha\beta}$, the second summand of the above formula is symmetric in (α, β). Thus $\int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^\beta_{t,k}\big)$ is symmetric in (α, β). □

Remark 2.7. It is easy to see that $\frac{\mathrm{i}}{2}\int_{X_0} \theta^\alpha \wedge \mathbb{H}\big(\mu(t)\lrcorner\eta^\beta_{t,N}\big) = M_{\alpha\gamma} A_{N+1}(t)^\beta_\gamma$. Thus $MA^T_{N+1}(t)$ is a symmetric matrix for $N \ge 0$.

3. A global result on deformation of holomorphic one-forms

This section will present a global convergence of the deformation of holomorphic one-forms in L^2 norm as a result of the following quasi-isometry for the operator $\bar\partial^* \circ \mathbb{G} \circ \partial$.

Proposition 3.1 [Liu et al. 2012a, Theorem 2.2.(3)]. *Let $T^{p,q} = \partial T^{p-1,q} \in A^{p,q}(M)$ on a compact Kähler manifold M. Then we have the inequality*

$$\big\|\bar\partial^* \circ \mathbb{G} \circ \partial T^{p-1,q}\big\|_{L^2} \le \|T^{p-1,q}\|_{L^2}. \tag{3-1}$$

Furthermore, if $T^{p-1,q}$ is ∂^*-exact, then the equality in (3-1) holds, i.e.,

$$\big\|\bar\partial^* \circ \mathbb{G} \circ \partial T^{p-1,q}\big\|_{L^2} = \|T^{p-1,q}\|_{L^2}.$$

This proposition was originally proved by step-by-step spectral decompositions in the preliminary version of [Liu et al. 2012a]. It is motivated by an attempt to prove the global Torelli theorem for the Teichmüller space of CY manifolds and inspired by the integral operators P and T defined by

$$Ph(\zeta) = -\frac{1}{\pi} \iint_{\mathbb{C}} h(z)\Big(\frac{1}{z-\zeta} - \frac{1}{z}\Big)\, dx\, dy, \quad \text{for functions } h \in L^p,\ p > 2,$$

$$Th(\zeta) = \lim_{\varepsilon\to 0} -\frac{1}{\pi} \iint_{|z-\zeta|>\varepsilon} \frac{h(z)}{(z-\zeta)^2}\, dx\, dy, \quad \text{for functions } h \in C^2_0.$$

These integral operators satisfy $(Ph)_{\bar z} = h$, $(Ph)_z = Th$, and the (quasi-)isometry relations

$$\|Th\|_{L^2} = \|h\|_{L^2},\ \ \|Th\|_{L^p} \le C_p\|h\|_{L^p},\ \text{ for any } p > 1, \text{ with } C_p \to 1 \text{ for } p \to 2,$$

which all appear in the classical Teichmüller theory for Riemann surfaces; see [Ahlfors 1966], whose main result is the proof of the existence of a solution of the Beltrami differential equation in $\mathbb{C}$

$$\frac{\partial}{\partial \bar{z}} w(\tau; z, \bar{z}) = \tau \mu (z, \bar{z}) \frac{\partial}{\partial z} w(\tau; z, \bar{z}),$$

where $\mu (z, \bar{z})$ is a Beltrami differential with $\|\mu (z, \bar{z})\|_\infty \leq c < 1$. The solution of the Beltrami differential equation is based on an iteration method due to Bojarski [1955], while it was Morrey [1938] who first proved the existence of the solution of the Beltrami equation. One of the main ingredients in the proof of the convergence of the Bojarski iteration method is the L^2-isometry of the inverse $\bar{\partial}^{-1}$ of the $\bar{\partial}$ operator in one complex variable. Kuranishi generalized the iteration method of Bojarski and constructed the Kuranishi map $\kappa : \mathbb{H}^{0,1}(X, T_X^{1,0}) \to \mathbb{H}^{0,2}(X, T_X^{1,0})$, the most basic technical tool in various aspects of deformation theory.

Fix a basis $\{\mu_i\}_{i=1}^n$ of harmonic $T_{X_p}^{(1,0)}$-valued (0,1) forms with

$$\sum_{i=1}^n \|\mu_i\|_{L^\infty} \leq 1,$$

and let $\mu(t) = \sum_{i=1}^n t_i \mu_i$ be the integrable Beltrami differential of the Kuranishi family $\varpi : \mathscr{X} \to \Delta_{p,1}$.

Theorem 3.2. *The* $(1, 0)$ *form* $\eta(t)$ *on* X_p *constructed in Theorem 2.1 converges in* L^2*-norm as long as* $|t| < 1$.

Proof. Recall that

$$\eta(t) = \theta + \sum_{i=1}^n t_i \eta_i + \sum_{|I|\geq 2} t^I \eta_I$$

is constructed as

$$\begin{cases} \eta_i = -\mathbb{G}\bar{\partial}^* \partial(\mu_i \lrcorner \theta), \\ \eta_{(i_1,\ldots,i_n)} = -\mathbb{G}\bar{\partial}^* \partial\big(\sum_{k=1}^n \mu_k \lrcorner \eta_{(i_1,\ldots,i_k-1,\ldots,i_n)}\big). \end{cases} \tag{3-2}$$

Here we identify η_i with $\eta_{(0,\ldots,\underline{1}_{i\text{-th}},\ldots,0)}$. Now let us discuss the global convergence in L^2-norm of the power series. By the quasi-isometry result in Proposition 3.1, together with (3-2) and the assumption $\sum_{i=1}^n \|\mu_i\|_{L^\infty} \leq 1$, we have

$$\sum_{|I|=i} \|\eta_I\|_{L^2} \leq \bigg(\sum_{i=1}^n \|\mu_i\|_{L^\infty}\bigg)\bigg(\sum_{|I|=i-1} \|\eta_I\|_{L^2}\bigg).$$

Also when $|I| = 1$, it is clear that

$$\sum_{i=1}^n \|\eta_i\|_{L^2} \leq \bigg(\sum_{i=1}^n \|\mu_i\|_{L^\infty}\bigg)\|\theta\|_{L^2} \leq \|\theta\|_{L^2}.$$

By induction, this yields that for every $k \geq 1$,

$$\sum_{|I|=k} \|\eta_I\|_{L^2} \leq \|\theta\|_{L^2},$$

which implies the estimates of $\eta(t)$

$$\|\eta(t)\|_{L^2} \leq \|\theta\|_{L^2} + \|\theta\|_{L^2} \sum_{\|I\|\geq 1} |t|^{|I|}. \qquad \square$$

Corollary 3.3. *The function $\theta(t)$ constructed in Theorem 2.1 converges in L^2-norm for $|t| < 1$.*

Proof. This follows directly from Theorem 3.2. $\square$

4. The Siegel metric and its curvature

Let us recall the local and global Torelli theorems of the moduli spaces of compact Riemann surfaces with genus g. Denote the Teichmüller space of the compact Riemann surface of genus g by $\mathcal{T}_g$ and the generalized Siegel upper half-plane by $\mathcal{H}_g$, which is $\{Z \in \mathrm{M}(g, \mathbb{C}) \mid Z = Z^t, \operatorname{Im} Z > 0\}$, an irreducible noncompact symmetric space, giving $\mathcal{A}_g$ a locally symmetric structure. Then

$$\Gamma_g(n) := \mathrm{Kernel}\big(\Gamma_g \xrightarrow{\rho} \mathrm{Sp}(g, \mathbb{Z}) \xrightarrow{\pi} \mathrm{Sp}(g, \mathbb{Z}/n\mathbb{Z})\big)$$

for $n \geq 2$ and $T_g := \mathrm{Kernel}\big(\Gamma_g \xrightarrow{\rho} \mathrm{Sp}(g, \mathbb{Z})\big)$, where Γ_g is the mapping class group of the compact Riemann surface of genus g. Also, the moduli space $\mathcal{M}_g^{(n)}$ of the compact Riemann surface of genus g with a fixed n-level structure is defined as the quotient of $\mathcal{T}_g$ by $\Gamma_g(n)$. We will fix $n \geq 3$ from now on. Meanwhile, the Torelli space $\mathcal{T}or_g$ is the quotient of $\mathcal{T}_g$ by T_g, called the Torelli group. Then we have the commutative diagram

$$\begin{tikzcd}
\mathcal{T}_g \arrow[d] \arrow[dr, "\Pi"] & \\
\mathcal{T}or_g \arrow[r, "\mathcal{J}^{\mathrm{tor}}"] \arrow[d] & \mathcal{H}_g \arrow[d] \\
\mathcal{M}_g^{(n)} \arrow[r, "\mathcal{J}^{(n)}"] \arrow[d] & \mathcal{A}_g^{(n)} \arrow[d] \\
\mathcal{M}_g \arrow[r, "\mathcal{J}"] & \mathcal{A}_g
\end{tikzcd}$$

$\mathcal{J}$ is always injective for $g \geq 2$. $\mathcal{J}^{\mathrm{tor}}$ is an open embedding for $g = 2$, while $\mathcal{J}^{\mathrm{tor}}$ and $\mathcal{J}^{(n)}$ are $2:1$ branched coverings onto its image ramified over hyperelliptic locus for $g \geq 3$. In other words, $\mathscr{J}^{\mathrm{tor}} : \mathscr{T}_g/\widetilde{T}_g \to \mathscr{H}_g$ is an embedding where $\widetilde{T}_g$ is defined as $\rho^{-1}(\langle -I_{2g}\rangle)$ and $\langle -I_{2g}\rangle$ is a subgroup of $\mathrm{Sp}(g, \mathbb{Z})$ generated by $-I_{2g}$. We shift to the local point of view. Π is everywhere an immersion for $g = 2$, but for the case $g \geq 3$, the tangent map of Π is injective on the nonhyperelliptic locus and vanishes on the normal directions of the hyperelliptic locus $\mathscr{HET}_g$. When restricted to $\mathscr{HET}_g$, Π is an immersion. According to [Liu et al. 2012b], the tangent map of $\mathscr{J}^{\mathrm{tor}} : \mathscr{T}_g/\widetilde{T}_g \to \mathscr{H}_g$ at the hyperelliptic locus from the Zariski tangent space of $\mathscr{T}_g/\widetilde{T}_g$ to the tangent space of $\mathscr{H}_g$ is injective.

Denote the Hodge bundle on $\mathscr{M}_g$ and $\mathscr{A}_g$ by $\mathscr{E}_g$; its fiber at a point is the vector space of holomorphic one-forms on $[X]$, a representative of the complex structure given by that point. There are three canonical metrics on $\mathscr{H}_g$ and $\mathscr{A}_g$, namely the Hodge metric, the Bergman metric and the Siegel metric. Hard Lefschetz decomposition and Hodge polarization give us a hermitian metric on $\mathscr{E}_g$, denoted by $\langle\,,\rangle$. From the natural isomorphism $\Omega^1_{\mathscr{H}_g} \cong S^2\mathscr{E}_g$, where S is the symmetric operator, it induces a hermitian metric on $T^{(1,0)}_{\mathscr{H}_g}$, denoted by $\tilde{\omega}_h$. The Bergman metric is defined by the Bergman kernel

$$\rho = -\log\det(\mathbb{1}_g - \overline{W^t}W) = -\log\det\left(\mathbb{1}_g - \overline{\left(\frac{\mathbb{1}_g + iZ}{\mathbb{1}_g - iZ}\right)^t}\left(\frac{\mathbb{1}_g + iZ}{\mathbb{1}_g - iZ}\right)\right),$$

where $W \in \left\{A \,\middle|\, A \in \mathrm{M}(g, \mathbb{C}),\, A^t = A,\, \mathbb{1}_g - \overline{A^t}A > 0\right\}$, which is the bounded domain, and Z is the coordinate of the Siegel upper half-plane

$$\left\{Z \,\middle|\, Z \in \mathrm{M}(g, \mathbb{C}),\, Z = Z^t,\, \mathrm{Im}(Z) > 0\right\}.$$

Here $\mathrm{M}(g, \mathbb{C})$ denotes the group of complex $g \times g$ matrices. Thus $\tilde{\omega}_b = \frac{\mathrm{i}}{2}\partial\bar{\partial}\rho$. Finally, the Siegel metric $\tilde{\omega}_s$ is defined by $\pi c_1(\mathscr{E}_g, \langle\,,\rangle)$. Pulled back by the period map, Siegel metrics, denoted by ω_s, also exist on $\mathscr{T}_g$, $\mathscr{T}_g$ and $\mathscr{M}_g$.

These three metrics are Kähler metrics and also invariant metrics on the irreducible homogeneous and symmetric space $\mathscr{H}_g$. It is clear that they are different by a constant multiple, while by [Yin 2010, Theorem 3.1 of Chapter 4], we know they are actually the same on $\mathscr{H}_g$.

Definition 4.1 (symmetric derivatives $\boldsymbol{S}_i$, $\boldsymbol{S}_{\bar{j}}$, $\boldsymbol{S}_{i\bar{j}}$, $\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}$ and $\boldsymbol{S}'_{i\bar{j}}\boldsymbol{S}'_{k\bar{l}}$). We give some examples to explain the use of these symbols. Here we use the notation $A := A(t)$, and similarly for $B(t)$, $C(t)$ and $D(t)$.

(1) First derivative: $\boldsymbol{S}_i$, $\boldsymbol{S}_{\bar{j}}$ and $\boldsymbol{S}_{i\bar{j}}$.

$$\boldsymbol{S}_i(A\bar{B}C\bar{D}) := \frac{\partial A}{\partial t_i}\bar{B}C\bar{D} + A\bar{B}\frac{\partial C}{\partial t_i}\bar{D}, \quad \boldsymbol{S}_{\bar{j}}(A\bar{B}C\bar{D}) := A\overline{\frac{\partial B}{\partial t_j}}C\bar{D} + A\bar{B}C\overline{\frac{\partial D}{\partial t_j}},$$

$$\boldsymbol{S}_{i\bar{j}}(A\bar{B}C\bar{D}) := \frac{\partial A}{\partial t_i}\overline{\frac{\partial B}{\partial t_j}}C\bar{D} + \frac{\partial A}{\partial t_i}\bar{B}C\overline{\frac{\partial D}{\partial t_j}} + A\overline{\frac{\partial B}{\partial t_j}}\frac{\partial C}{\partial t_i}\bar{D} + A\bar{B}\frac{\partial C}{\partial t_i}\overline{\frac{\partial D}{\partial t_j}},$$

where $A, B, C, D \in \mathrm{M}(n, \mathbb{C})$ are all holomorphic in t. Indices without a bar mean taking derivatives through all holomorphic matrices, and indices with a bar do so through all antiholomorphic matrices.

(2) Second derivative: $\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}$ and $\boldsymbol{S}'_{i\bar{j}}\boldsymbol{S}'_{k\bar{l}}$.

$\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}(A\overline{B}C\overline{D})$

$$\begin{aligned}
:=&\frac{\partial^2 A}{\partial t_i\partial t_k}\overline{\frac{\partial^2 B}{\partial t_j\partial t_l}}C\overline{D}+\frac{\partial A}{\partial t_i}\overline{\frac{\partial^2 B}{\partial t_j\partial t_l}}\frac{\partial C}{\partial t_k}\overline{D}+\frac{\partial A}{\partial t_k}\overline{\frac{\partial^2 B}{\partial t_j\partial t_l}}\frac{\partial C}{\partial t_i}\overline{D}+A\overline{\frac{\partial^2 B}{\partial t_j\partial t_l}}\frac{\partial^2 C}{\partial t_i\partial t_k}\overline{D}\\
&+\frac{\partial^2 A}{\partial t_i\partial t_k}\overline{\frac{\partial B}{\partial t_j}}C\overline{\frac{\partial D}{\partial t_l}}+\frac{\partial A}{\partial t_i}\overline{\frac{\partial B}{\partial t_j}}\frac{\partial C}{\partial t_k}\overline{\frac{\partial D}{\partial t_l}}+\frac{\partial A}{\partial t_k}\overline{\frac{\partial B}{\partial t_j}}\frac{\partial C}{\partial t_i}\overline{\frac{\partial D}{\partial t_l}}+A\overline{\frac{\partial B}{\partial t_j}}\frac{\partial^2 C}{\partial t_i\partial t_k}\overline{\frac{\partial D}{\partial t_l}}\\
&+\frac{\partial^2 A}{\partial t_i\partial t_k}\overline{\frac{\partial B}{\partial t_l}}C\overline{\frac{\partial D}{\partial t_j}}+\frac{\partial A}{\partial t_i}\overline{\frac{\partial B}{\partial t_l}}\frac{\partial C}{\partial t_k}\overline{\frac{\partial D}{\partial t_j}}+\frac{\partial A}{\partial t_k}\overline{\frac{\partial B}{\partial t_l}}\frac{\partial C}{\partial t_i}\overline{\frac{\partial D}{\partial t_j}}+A\overline{\frac{\partial B}{\partial t_l}}\frac{\partial^2 C}{\partial t_i\partial t_k}\overline{\frac{\partial D}{\partial t_j}}\\
&+\frac{\partial^2 A}{\partial t_i\partial t_k}\overline{B}C\overline{\frac{\partial^2 D}{\partial t_j\partial t_l}}+\frac{\partial A}{\partial t_i}\overline{B}\frac{\partial C}{\partial t_k}\overline{\frac{\partial^2 D}{\partial t_j\partial t_l}}+\frac{\partial A}{\partial t_k}\overline{B}\frac{\partial C}{\partial t_i}\overline{\frac{\partial^2 D}{\partial t_j\partial t_l}}+A\overline{B}\frac{\partial^2 C}{\partial t_i\partial t_k}\overline{\frac{\partial^2 D}{\partial t_j\partial t_l}},
\end{aligned}$$

$\boldsymbol{S}'_{i\bar{j}}\boldsymbol{S}'_{k\bar{l}}(A\overline{B}C\overline{D})$

$$:=\frac{\partial A}{\partial t_i}\overline{\frac{\partial B}{\partial t_j}}\frac{\partial C}{\partial t_k}\overline{\frac{\partial D}{\partial t_l}}+\frac{\partial A}{\partial t_k}\overline{\frac{\partial B}{\partial t_j}}\frac{\partial C}{\partial t_i}\overline{\frac{\partial D}{\partial t_l}}+\frac{\partial A}{\partial t_i}\overline{\frac{\partial B}{\partial t_l}}\frac{\partial C}{\partial t_k}\overline{\frac{\partial D}{\partial t_j}}+\frac{\partial A}{\partial t_i}\overline{\frac{\partial B}{\partial t_l}}\frac{\partial C}{\partial t_k}\overline{\frac{\partial D}{\partial t_j}}.$$

The difference between these two symbols lies in that $\dfrac{\partial}{\partial t_i}$ and $\dfrac{\partial}{\partial t_k}$ can't operate on a matrix simultaneously in $\boldsymbol{S}'_{i\bar{j}}\boldsymbol{S}'_{k\bar{l}}$.

Theorem 4.2. *The Siegel metric $\omega_s(t)$ on the nonhyperelliptic locus of $\mathcal{T}_g$ can be written as*

$$\omega_s(t)=\frac{\mathrm{i}}{2}\sum_{n=1}^{\infty}\frac{1}{n}\partial\bar{\partial}\,\mathrm{tr}(A(t)\overline{A(t)})^n.$$

Proof. From the definition of the Siegel metric and the fact that holomorphic one-forms on the Riemann surface and its Jacobian torus can be identified, we will write it out explicitly using the Kuranishi coordinate $\Delta_{p,\epsilon}$ on $\mathcal{T}_g$ with $g\geq 3$ and the deformation formula for holomorphic one-forms, where p lies in the nonhyperelliptic locus.

$$\begin{aligned}
\text{(4-1)}\quad \omega_s(t)&=\pi c_1(\mathscr{E}_g,\langle\ \rangle)\\
&=-\frac{\mathrm{i}}{2}\partial\bar{\partial}\log\det\frac{\mathrm{i}}{2}\int_{X_p}\theta^{\alpha}(t)\wedge\overline{\theta^{\beta}(t)}\\
&=-\frac{\mathrm{i}}{2}\partial\bar{\partial}\,\mathrm{tr}\log\frac{\mathrm{i}}{2}\int_{X_p}\big(\theta_p^{\alpha}+A(t)^{\alpha}_{\gamma}\bar{\theta}_p^{\gamma}+E^{\alpha}(t)\big)\wedge\overline{\big(\theta_p^{\beta}+A(t)^{\beta}_{\delta}\bar{\theta}_p^{\delta}+E^{\beta}(t)\big)}\\
&=-\frac{\mathrm{i}}{2}\partial\bar{\partial}\,\mathrm{tr}\log\big(M_{p,\alpha\beta}-A(t)^{\alpha}_{\gamma}\overline{A(t)^{\beta}_{\delta}}M_{p,\delta\gamma}\big)
\end{aligned}$$

$$= -\frac{\mathrm{i}}{2}\partial\bar{\partial}\operatorname{tr}\log M_{p,\alpha\eta}\big(\delta_{\eta\beta} - M_p^{\eta\sigma}A(t)_\gamma^\sigma\overline{A(t)_\delta^\beta}M_{p,\delta\gamma}\big)$$
$$= -\frac{\mathrm{i}}{2}\partial\bar{\partial}\operatorname{tr}\log\big(\delta_{\alpha\beta} - M_p^{\alpha\sigma}A(t)_\gamma^\sigma\overline{A(t)_\delta^\beta}M_{p,\delta\gamma}\big).$$

We remark here that the $\frac{\mathrm{i}}{2}\int_{X_p}\theta^\alpha(t)\wedge\overline{\theta^\beta(t)}$ are positive hermitian matrices for t small, and thus diagonalizable matrices. Thus it makes sense for the operator $\operatorname{tr}\log$.

Formulate all these into the matrix type to get

$$\omega_s(t) = -\frac{\mathrm{i}}{2}\partial\bar{\partial}\operatorname{tr}\log\big(I - M_p^{-1}A(t)M_p\overline{A(t)}^T\big).$$

From the Hodge–Riemann bilinear relation $A(t)M_p = M_pA(t)^T$, it follows that

$$M_p^{-1}A(t)M_p\overline{A(t)}^T = A(t)^T\overline{A(t)}^T.$$

Then the Siegel metric $\omega_s(t)$ is given by

$$\omega_s(t) = -\frac{\mathrm{i}}{2}\partial\bar{\partial}\operatorname{tr}\log\big(\, I - A(t)^T\overline{A(t)}^T\big) = \frac{\mathrm{i}}{2}\partial\bar{\partial}\operatorname{tr}\sum_{n=1}^{\infty}\frac{1}{n}\big(A(t)^T\overline{A(t)}^T\big)^n \tag{4-2}$$
$$= \frac{\mathrm{i}}{2}\sum_{n=1}^{\infty}\frac{1}{n}\partial\bar{\partial}\operatorname{tr}\big(A(t)^T\overline{A(t)}^T\big)^n = \frac{\mathrm{i}}{2}\sum_{n=1}^{\infty}\frac{1}{n}\partial\bar{\partial}\operatorname{tr}\big(A(t)\overline{A(t)}\big)^n.$$

Restricted to the origin, the Siegel metric is $\omega_s(0) = \frac{\mathrm{i}}{2}\sum_{i,j=1}^{n}\operatorname{tr}(A_i\overline{A_j})dt_i\wedge d\bar{t}_j$. □

To compute the curvature of the Siegel metric, we rewrite (3-2) according to the degree of t:

$$\omega_s(t) = \frac{\mathrm{i}}{2}\partial\bar{\partial}\sum_{k\geq 2}\sum_{\substack{m_i>0,1\leq i\leq 2l\\ m_1+\cdots+m_{2l}=k}}\frac{1}{l}\operatorname{tr}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\ldots\overline{A_{m_{2l}}(t)}\big) \tag{4-3}$$
$$= \frac{\mathrm{i}}{2}\sum_{k\geq 2}\sum_{\substack{m_i>0,1\leq i\leq 2l\\ m_1+\cdots+m_{2l}=k}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{i\bar{j}}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\ldots\overline{A_{m_{2l}}(t)}\big)\big)dt_i\wedge d\bar{t}_j$$
$$= \frac{\mathrm{i}}{2}\sum_{k\geq 0}\sum_{\substack{m_i>0,1\leq i\leq 2l\\ m_1+\cdots+m_{2l}=k+2}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{i\bar{j}}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\ldots\overline{A_{m_{2l}}(t)}\big)\big)dt_i\wedge d\bar{t}_j.$$

From (4-3), we know that, if we set $\omega_s(t) = \frac{\mathrm{i}}{2}\omega_{i\bar{j}}dt_i\wedge d\bar{t}_j$, then

$$\omega_{i\bar{j}} = \sum_{k\geq 0}\sum_{\substack{m_i>0,1\leq i\leq 2l\\ m_1+\cdots+m_{2l}=k+2}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{i\bar{j}}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\ldots\overline{A_{m_{2l}}(t)}\big)\big), \tag{4-4}$$

with $\omega_{i\bar{j}}(0) = \operatorname{tr}(A_i\overline{A_j})$.

We need an auxiliary combinatorial lemma before getting to the curvature formula.

Lemma 4.3. *The function $h_{ij}(t)$ has the expansion*

$$\delta_{ij} + (h_{ij})_1 + (h_{ij})_2 + \cdots,$$

where $(h_{ij})_n$ is the n-th order part of the expansion; *then $h^{ij}(t)$ can be expanded as*

$$h^{ij}(t) = \delta_{ij} - (h_{ij})_1 + \sum_{k\geq 2} \sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=k}} (-1)^l (h_{ii_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} j})_{m_l}.$$

Proof. Directly check that $h_{ij}(t)h^{jk}(t) = \delta_{ik}$, which is equivalent to

$$\begin{aligned}\delta_{ik} = &\Bigl[\delta_{ij} + (h_{ij})_1 + \sum_{p\geq 2}(h_{ij})_p\Bigr]\\ &\times\Bigl[\delta_{jk} - (h_{jk})_1 + \sum_{p\geq 2} \sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=p}} (-1)^l (h_{ji_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} k})_{m_l}\Bigr].\end{aligned}$$

It is quite easy to see that the zeroth- and first-order parts of both sides coincide. Thus this reduces to checking that for $p \geq 2$,

$$\begin{aligned}0 = (h_{ik})_p - (h_{ij})_{p-1}(h_{jk})_1 + \cdots &\\ + (h_{ij})_1 &\sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=p-1}} (-1)^l (h_{ji_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} k})_{m_l}\\ + &\sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=p}} (-1)^l (h_{ii_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} k})_{m_l}.\end{aligned}$$

The right-hand side can be written as

$$\begin{aligned}&\sum_{i=1}^{p} \sum_{\substack{m_1=i, m_j>0, 2\leq j\leq l\\ m_1+\cdots+m_l=p}} (-1)^{l-1} (h_{ii_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} k})_{m_l}\\ &\qquad + \sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=p}} (-1)^l (h_{ii_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} k})_{m_l}\\ &= \sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=p}} (-1)^{l-1} (h_{ii_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} k})_{m_l}\\ &\qquad + \sum_{\substack{m_i>0, 1\leq i\leq l\\ m_1+\cdots+m_l=p}} (-1)^l (h_{ii_1})_{m_1} (h_{i_1 i_2})_{m_2} \ldots (h_{i_{l-1} k})_{m_l}.\end{aligned}$$

Now clearly this is zero. Our lemma is proved. □

Choose a normal coordinate around p such that $\omega_{i\bar{j}}(0) = \delta_{ij}$, $(\partial\omega_{i\bar{j}}/\partial t_k)(0) = (\partial\omega_{i\bar{j}}/\partial \bar{t}_k)(0) = 0$, and $(\partial\omega_{i\bar{j}}/\partial t_k \partial t_l)(0) = (\partial\omega_{i\bar{j}}/\partial \bar{t}_k \partial \bar{t}_l)(0) = 0$, still denoted by $\Delta_{p,\epsilon}$. According to the convention of $A_N(t)$ we make after the definition of $A(t)$ and $E(t)$, this is equivalent to saying

$$\begin{cases} \operatorname{tr}(A_i \overline{A_j}) = \delta_{ij}, \\ \operatorname{tr}(A_i \overline{A_{jk}}) = \operatorname{tr}(A_{ik}\overline{A_j}) = 0, \\ \operatorname{tr}(A_{ikl}\overline{A_j}) = \operatorname{tr}(A_i \overline{A_{jkl}}) = 0. \end{cases} \tag{4-5}$$

From Lemma 4.3, we get

$$\begin{aligned} \omega^{\bar{i}j} = \delta_{ij} + \sum_{k\geq 1} \sum_{\substack{m_i>0, 1\leq i\leq l \\ m_1+\cdots+m_l=k}} (-1)^l \prod_{i=1}^{l} \frac{1}{s_i} & \\ \times \Biggl(\sum_{\substack{m_{1n}>0, 1\leq n\leq 2s_1 \\ \sum_{n=1}^{2s_1} m_{1n}=m_1+2}} & \operatorname{tr}\bigl(\boldsymbol{S}_{i\bar{i}_1}\bigl(A_{m_{11}}(t)\dots\overline{A_{m_{12s_1}}(t)}\bigr)\bigr)\Biggr) \\ \times \Biggl(\sum_{\substack{m_{2n}>0, 1\leq n\leq 2s_2 \\ \sum_{n=1}^{2s_2} m_{2n}=m_2+2}} & \operatorname{tr}\bigl(\boldsymbol{S}_{i_1\bar{i}_2}\bigl(A_{m_{21}}(t)\dots\overline{A_{m_{22s_2}}(t)}\bigr)\bigr)\Biggr) \\ \times\cdots\times \Biggl(\sum_{\substack{m_{ln}>0, 1\leq n\leq 2s_l \\ \sum_{n=1}^{2s_l} m_{ln}=m_l+2}} & \operatorname{tr}\bigl(\boldsymbol{S}_{i_{l-1}\bar{j}}\bigl(A_{m_{l1}}(t)\dots\overline{A_{m_{l2s_l}}(t)}\bigr)\bigr)\Biggr). \end{aligned} \tag{4-6}$$

Theorem 4.4. *The curvature* $\mathrm{R}_{i\bar{j}k\bar{l}}$ *of the Siegel metric* $\omega_s(t)$ *is given by*

$$\begin{aligned} \mathrm{R}_{i\bar{j}k\bar{l}} = -\sum_{N\geq 0} \sum_{\substack{m_i>0, 1\leq i\leq 2l \\ m_1+\cdots+m_{2l}=N+4}} \frac{1}{l} \operatorname{tr}\bigl(\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}\bigl(A_{m_1}(t)\overline{A_{m_2}(t)}\dots\overline{A_{m_{2l}}(t)}\bigr)\bigr) & \\ + \sum_{N\geq 0} \sum_{\substack{N_i\geq 0, 1\leq i\leq 3 \\ \sum_{i=1}^{3} N_i=N}} \Biggl[\sum_{\substack{m_i>0, 1\leq i\leq l \\ m_1+\cdots+m_l=N_1}} (-1)^l \prod_{i=1}^{l} \frac{1}{s_i} & \\ \times \sum_{\substack{m_{1n}>0, 1\leq n\leq 2s_1 \\ \sum_{n=1}^{2s_1} m_{1n}=m_1+2}} \operatorname{tr}\bigl(\boldsymbol{S}_{q\bar{i}_1}\bigl(A_{m_{11}}(t)\cdots\overline{A_{m_{12s_1}}(t)}\bigr)\bigr) & \\ \times \sum_{\substack{m_{2n}>0, \\ \sum_{n=1}^{2s_2} m_{2n}=m_2+2}} \operatorname{tr}\bigl(\boldsymbol{S}_{i_1\bar{i}_2}\bigl(A_{m_{21}}(t)\dots\overline{A_{m_{22s_2}}(t)}\bigr)\bigr)\dots & \\ \times \sum_{\substack{m_{ln}>0, \\ \sum_{n=1}^{2s_l} m_{ln}=m_l+2}} \operatorname{tr}\bigl(\boldsymbol{S}_{i_{l-1}\bar{p}}\bigl(A_{m_{l1}}(t)\dots\overline{A_{m_{l2s_l}}(t)}\bigr)\bigr)\Biggr] & \end{aligned}$$

$$\times\Bigg[\sum_{\substack{m_i>0,\\ \sum_{i=1}^{2l} m_i=N_2+3}} \frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_i\boldsymbol{S}_{k\bar q}\big(A_{m_1}(t)\dots\overline{A_{m_{2l}}(t)}\big)\big)\Bigg]$$
$$\times\Bigg[\sum_{\substack{m_i>0,\\ \sum_{i=1}^{2l} m_i=N_3+3}} \frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{\bar j}\boldsymbol{S}_{p\bar l}\big(A_{m_1}(t)\dots\overline{A_{m_{2l}}(t)}\big)\big)\Bigg],$$

where we need the convention that the first square bracket in the second summand will be δ_{qp} as $N_1=0$.

Proof. Just use the well known curvature formula

$$\mathrm{R}_{i\bar j k\bar l}=-\frac{\partial^2\omega_{k\bar l}}{\partial t_i\partial\bar t_j}+\omega^{\bar q p}\frac{\partial\omega_{k\bar q}}{\partial t_i}\frac{\partial\omega_{p\bar l}}{\partial\bar t_j}.$$

By use of (4-4), we have

$$\frac{\partial^2\omega_{k\bar l}}{\partial t_i\partial\bar t_j}=\sum_{n\ge 0}\sum_{\substack{m_i>0,1\le i\le 2l\\ m_1+\cdots+m_{2l}=n+4}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{i\bar j}\boldsymbol{S}_{k\bar l}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\dots\overline{A_{m_{2l}}(t)}\big)\big),$$
$$\frac{\partial\omega_{k\bar q}}{\partial t_i}=\sum_{n\ge 0}\sum_{\substack{m_i>0,1\le i\le 2l\\ m_1+\cdots+m_{2l}=n+3}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_i\boldsymbol{S}_{k\bar q}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\dots\overline{A_{m_{2l}}(t)}\big)\big),$$
$$\frac{\partial\omega_{p\bar l}}{\partial\bar t_j}=\sum_{n\ge 0}\sum_{\substack{m_i>0,1\le i\le 2l\\ m_1+\cdots+m_{2l}=n+3}}\frac{1}{l}\operatorname{tr}\big(\boldsymbol{S}_{\bar j}\boldsymbol{S}_{p\bar l}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\dots\overline{A_{m_{2l}}(t)}\big)\big).$$

From this we get the formula (4-6) of $\omega^{\bar i j}$. Thus the theorem follows easily. □

Also the curvature of the Siegel metric at the point p can be easily gotten from the curvature formula together with (4-5):

$$\begin{aligned}\text{(4-7)}\quad \mathrm{R}_{i\bar j k\bar l}(0)&=-\operatorname{tr}\big(\boldsymbol{S}_{i\bar j}\boldsymbol{S}_{k\bar l}\big(A_2(t)\overline{A_2(t)}\big)\big)-\tfrac12\operatorname{tr}\big(\boldsymbol{S}_{i\bar j}\boldsymbol{S}_{k\bar l}\big(A_1(t)\overline{A_1(t)}A_1(t)\overline{A_1(t)}\big)\big)\\&\quad+\operatorname{tr}\big(\boldsymbol{S}_i\boldsymbol{S}_{k\bar p}\big(A_2(t)\overline{A_1(t)}\big)\big)\operatorname{tr}\big(\boldsymbol{S}_{\bar j}\boldsymbol{S}_{p\bar l}\big(A_1(t)\overline{A_2(t)}\big)\big)\\&=-\operatorname{tr}(A_{ik}\overline{A_{jl}})-\operatorname{tr}(A_i\overline{A_j}A_k\overline{A_l})\\&\quad-\operatorname{tr}(A_i\overline{A_l}A_k\overline{A_j})+\operatorname{tr}(A_{ik}\overline{A_p})\operatorname{tr}(A_p\overline{A_{jl}})\\&=-\operatorname{tr}(A_{ik}\overline{A_{jl}})-\operatorname{tr}(A_i\overline{A_j}A_k\overline{A_l})-\operatorname{tr}(A_i\overline{A_l}A_k\overline{A_j}).\end{aligned}$$

The holomorphic sectional curvature along the direction $v=\sum_{i=1}^{3g-3}a_i\mu_i\in\mathbb{H}^{0,1}_{\bar\partial}(X_p,T_{X_p})$ is given by

$$\text{(4-8)} \qquad \mathrm{H}(v) = \frac{\mathrm{R}_{v\bar{v}v\bar{v}}}{|v|^4} = \frac{\sum_{i,j,k,l=1}^{3g-3} a_i \overline{a_j} a_k \overline{a_l} \mathrm{R}_{i\bar{j}k\bar{l}}(0)}{\left(\sum_{i,j=1}^{3g-3} a_i \overline{a_j} \omega_{i\bar{j}}(0)\right)^2}$$
$$= \frac{-2\,\mathrm{tr}\big((a_i A_i)\overline{(a_j A_j)}(a_k A_k)\overline{(a_l A_l)}\big) - \mathrm{tr}\big((a_i a_k A_{ik})(\overline{a_j a_l A_{jl}})\big)}{\left(a_i \overline{a_j}\,\mathrm{tr}(A_i \overline{A_j})\right)^2}$$
$$\le -2\frac{\mathrm{tr}\big((a_i A_i)\overline{(a_j A_j)}(a_k A_k)\overline{(a_l A_l)}\big)}{\left(\mathrm{tr}((a_i A_i)\overline{(a_j A_j)})\right)^2}.$$

Set $\sum_{i=1}^{3g-3} a_i A_i = E$ and normalize M_p to $\mathrm{i}\mathbb{1}_g$. Then $A(t)$ is symmetric and A_i are all symmetric for $1 \le i \le g$. By the mean value inequality,

$$\text{(4-9)} \qquad 1 \ge \frac{\mathrm{tr}(E\bar{E}E\bar{E})}{(\mathrm{tr}(E\bar{E}))^2} \ge \frac{1}{g}$$

for the symmetric matrix E. The proof of Proposition 5.4 contains further details. Thus we have

$$\mathrm{H}(v) \le -\frac{2}{g}.$$

5. The second fundamental form of a nonhyperelliptic locus and the totally geodesic submanifold

Now we are ready to compute the second fundamental form of $\mathcal{J} : \mathcal{M}_g \to \mathcal{A}_g$, always fixing the Siegel metric $\widetilde{\omega}_s$ on $\mathcal{A}_g$. Lift to $\mathcal{J}^{\mathrm{tor}} : \mathcal{T}_g \to \mathcal{H}_g$, with Siegel metric $\widetilde{\omega}_s$ on $\mathcal{H}_g$. The local Torelli theorem assures the exact sequence

$$0 \to T^{(1,0)}_{\mathcal{T}_g} \to \mathcal{J}^{\mathrm{tor}\,*} T^{(1,0)}_{\mathcal{H}_g} \xrightarrow{\pi} N \to 0$$

when restricted to a nonhyperelliptic locus of $\mathcal{T}_g$ and when N is the normal bundle. Also we have the natural connection $\mathcal{J}^{\mathrm{tor}\,*}\nabla$ on $\mathcal{J}^{\mathrm{tor}\,*}T^{(1,0)}_{\mathcal{H}_g}$, where the Chern connection ∇ is determined by $\widetilde{\omega}_s$ on $\mathcal{H}_g$. Following the argument of [Colombo and Frediani 2010, pp. 6–7], the second fundamental form σ is defined by

$$\sigma(s) = \pi(\nabla s),\ s \in A^0\big(T^{(1,0)}_{\mathcal{T}_g}\big).$$

From the Gauss equation, it follows that

$$\left(R\left(\frac{\partial}{\partial t_k}\right), \frac{\partial}{\partial t_l}\right)\left(\frac{\partial}{\partial t_i}, \frac{\partial}{\partial t_j}\right) = \left(\widetilde{R}\left(\frac{\partial}{\partial t_k}\right), \frac{\partial}{\partial t_l}\right)\left(\frac{\partial}{\partial t_i}, \frac{\partial}{\partial t_j}\right) - \left(\sigma\left(\frac{\partial}{\partial t_k}\right), \sigma\left(\frac{\partial}{\partial t_l}\right)\right)\left(\frac{\partial}{\partial t_i}, \frac{\partial}{\partial t_j}\right),$$

where R is the curvature operator of the Siegel metric on $\mathcal{T}_g$, while $\widetilde{R}$ is the one on $\mathscr{H}_g$. Set

$$\Sigma_{i\bar{j}k\bar{l}} = \left(\sigma\left(\frac{\partial}{\partial t_k}\right), \sigma\left(\frac{\partial}{\partial t_l}\right)\right)\left(\frac{\partial}{\partial t_i}, \frac{\partial}{\partial t_j}\right).$$

Thus

$$R_{i\bar{j}k\bar{l}} = \widetilde{R}_{i\bar{j}k\bar{l}} - \Sigma_{i\bar{j}k\bar{l}}. \tag{5-1}$$

Thus we focus on the Siegel metric $\widetilde{\omega}_s$ and its curvature, and use the bounded domain $\mathscr{B}_g$ to simplify the computation.

Siegel geometry on $\mathscr{B}_g$.

Theorem 5.1. *The Siegel metric $\widetilde{\omega}_s$ on $\mathscr{B}_g$ has the full expansion formula*

$$\widetilde{\omega}_s = \frac{\mathrm{i}}{2}\sum_{n\geq 0}\sum_{\alpha\leq\beta}\sum_{\gamma\leq\delta}\frac{1}{n+1}\operatorname{tr}\big(\boldsymbol{S}_{(\alpha\beta)\overline{(\gamma\delta)}}(W\overline{W})^{n+1}\big)dW_{\alpha\beta}\wedge d\overline{W_{\gamma\delta}}.$$

Proof. Because the Siegel metric equals the Bergman metric, we use the Bergman kernel

$$\begin{aligned}\widetilde{\omega}_s &= -\frac{\mathrm{i}}{2}\partial\bar{\partial}\log\det(\mathbb{1}_g - \overline{W}W) = -\frac{\mathrm{i}}{2}\partial\bar{\partial}\ \operatorname{tr}\log(\mathbb{1}_g - \overline{W}W)\\ &= \frac{\mathrm{i}}{2}\partial\bar{\partial}\left(\sum_{n=1}^{\infty}\frac{1}{n}\operatorname{tr}(\overline{W}W)^n\right) = \frac{\mathrm{i}}{2}\partial\bar{\partial}\left(\sum_{n=1}^{\infty}\frac{1}{n}\operatorname{tr}(W\overline{W})^n\right).\end{aligned}$$

Now $\widetilde{\omega}_s$ can be written as

$$\widetilde{\omega}_s = \frac{\mathrm{i}}{2}\sum_{n\geq 0}\sum_{\alpha\leq\beta}\sum_{\gamma\leq\delta}\frac{1}{n+1}\operatorname{tr}\big(\boldsymbol{S}_{(\alpha\beta)\overline{(\gamma\delta)}}(W\overline{W})^{n+1}\big)dW_{\alpha\beta}\wedge d\overline{W_{\gamma\delta}},$$

where $\boldsymbol{S}_{(\alpha\beta)\overline{(\gamma\delta)}}$ indicates taking derivatives along $\partial/\partial W_{\alpha\beta}$ and $\partial/\partial\overline{W_{\gamma\delta}}$ with $\alpha\leq\beta, \gamma\leq\delta$ according to Definition 4.1. Since W is symmetric, $\partial/\partial W_{\alpha\beta}$ takes the derivative with respect to $W_{\alpha\beta}$ and $W_{\beta\alpha}$. $\square$

Similarly, if we write $\widetilde{\omega}_s = \sum_{\alpha\leq\beta}\sum_{\gamma\leq\delta}\frac{\mathrm{i}}{2}\widetilde{\omega}_{(\alpha\beta)\overline{(\gamma\delta)}}dW_{\alpha\beta}\wedge d\overline{W_{\gamma\delta}}$, it is easy to see that

$$\widetilde{\omega}_{(\alpha\beta)\overline{(\gamma\delta)}}(0) = \begin{cases}1 & \text{for } \alpha=\gamma=\beta=\delta,\\ 2 & \text{for } \alpha=\gamma\neq\beta=\delta,\\ 0 & \text{otherwise,}\end{cases}$$

and

$$\widetilde{\omega}_{(\alpha\beta)\overline{(\gamma\delta)}} = \frac{\mathrm{i}}{2}\sum_{n\geq 0}\frac{1}{n+1}\operatorname{tr}\big(\boldsymbol{S}_{(\alpha\beta)\overline{(\gamma\delta)}}(W\overline{W})^{n+1}\big). \tag{5-2}$$

In the following computation, the matrix D is defined as

$$D_{\alpha\beta} = \begin{cases}1 & \text{for } \alpha=\beta,\\ \sqrt{2} & \text{for } \alpha\neq\beta.\end{cases}$$

Lemma 5.2. *We have*

$$
\begin{aligned}
\widetilde{\omega}^{\overline{(\alpha\beta)}(\gamma\delta)} &= \frac{1}{D_{\alpha\beta}^2}\frac{1}{D_{\gamma\delta}^2}\widetilde{\omega}_{(\alpha\beta)\overline{(\gamma\delta)}}(0) \\
&+\sum_{k\geq 1}\sum_{\substack{m_i>0,1\leq i\leq l\\ m_1+\cdots+m_l=k}}(-1)^l\frac{1}{D_{\alpha\beta}^2}\frac{1}{D_{\gamma\delta}^2}\biggl(\frac{1}{m_1+1}\operatorname{tr}\bigl(\boldsymbol{S}_{(\alpha\beta)\overline{(\alpha_1\beta_1)}}(W\overline{W})^{m_1+1}\bigr)\biggr) \\
&\times\biggl(\frac{1}{m_2+1}\operatorname{tr}\bigl(\boldsymbol{S}_{(\alpha_1\beta_1)\overline{(\alpha_2\beta_2)}}(W\overline{W})^{m_2+1}\bigr)\biggr)\cdots\biggl(\frac{1}{m_l+1}\operatorname{tr}\bigl(\boldsymbol{S}_{(\alpha_{l-1}\beta_{l-1})\overline{(\gamma\delta)}}(W\overline{W})^{m_l+1}\bigr)\biggr),
\end{aligned}
$$

where $\alpha_1, \beta_1, \ldots, \alpha_{l-1}, \beta_{l-1}$ are free indices from 1 *to* g. (Here $(\alpha_i\beta_i)$ means taking the derivative just with respect to $W_{\alpha_i\beta_i}$; this convention will remain in force later.)

Proof. We use another coordinate $Y_{\alpha\beta} := D_{\alpha\beta}X_{\alpha\beta}$ on $\mathfrak{B}_g$ to rewrite $\widetilde{\omega}_s$. It is easy to check that $\widetilde{\omega}_s = \frac{\mathrm{i}}{2}\sum_{\alpha\leq\beta}\sum_{\gamma\leq\delta}\widetilde{\Omega}_{(\alpha\beta)\overline{(\gamma\delta)}}\,dY_{\alpha\beta}\wedge d\overline{Y_{\gamma\delta}}$, where

$$
\widetilde{\Omega}_{(\alpha\beta)\overline{(\gamma\delta)}} = \frac{1}{D_{\alpha\beta}}\frac{1}{D_{\gamma\delta}}\widetilde{\omega}_{(\alpha\beta)\overline{(\gamma\delta)}} \tag{5-3}
$$

and $\widetilde{\Omega}_{(\alpha\beta)\overline{(\gamma\delta)}}(0) = \delta_{\alpha\gamma}\delta_{\delta\beta}$ (Kronecker symbol). Now an application of Lemma 4.3 to $\widetilde{\Omega}_{(\alpha\beta)\overline{(\gamma\delta)}}$ yields

$$
\begin{aligned}
\widetilde{\Omega}^{\overline{(\alpha\beta)}(\gamma\delta)} = \widetilde{\Omega}_{(\alpha\beta)\overline{(\gamma\delta)}}(0)+\sum_{k\geq 1}\sum_{\substack{m_i>0,1\leq i\leq n\\ m_1+\cdots+m_l=k}}\sum_{\alpha_1\leq\beta_1}\cdots\sum_{\alpha_{l-1}\leq\beta_{l-1}}(-1)^l\bigl(\widetilde{\Omega}_{(\alpha\beta)\overline{(\alpha_1\beta_1)}}\bigr)_{m_1}\cdots \\
\times\bigl(\widetilde{\Omega}_{(\alpha_{l-1}\beta_{l-1})\overline{(\gamma\delta)}}\bigr)_{m_l}.
\end{aligned}
$$

From (5-3) and the equality $\widetilde{\Omega}^{\overline{(\alpha\beta)}(\gamma\delta)} = D_{\alpha\beta}D_{\gamma\delta}\widetilde{\omega}^{\overline{(\alpha\beta)}(\gamma\delta)}$, we get the result. □

Theorem 5.3. *The curvature $\widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}}$ is given by*

$$
\begin{aligned}
\widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}} &= -\sum_{N\geq 0}\frac{1}{N+2}\operatorname{tr}\bigl(\boldsymbol{S}_{(\alpha\beta)\overline{(\gamma\delta)}}\boldsymbol{S}_{(\zeta\eta)\overline{(\sigma\tau)}}\bigl(W\overline{W}\bigr)^{N+2}\bigr) \\
&+\sum_{N\geq 0}\sum_{\substack{N_1\geq 0,N_2>0,N_3>0\\ \sum_{i=1}^3 N_i=N+1}}\Biggl[\sum_{\substack{m_i>0,1\leq i\leq l\\ m_1+\cdots+m_l=N_1}}(-1)^l\biggl(\frac{1}{m_1+1}\operatorname{tr}\bigl(\boldsymbol{S}_{(\lambda\mu)\overline{(\alpha_1\beta_1)}}(W\overline{W})^{m_1+1}\bigr)\biggr) \\
&\times\biggl(\frac{1}{m_2+1}\operatorname{tr}\bigl(\boldsymbol{S}_{(\alpha_1\beta_1)\overline{(\alpha_2\beta_2)}}(W\overline{W})^{m_2+1}\bigr)\biggr)\cdots\biggl(\frac{1}{m_l+1}\operatorname{tr}\bigl(\boldsymbol{S}_{(\alpha_{l-1}\beta_{l-1})\overline{(\phi\psi)}}(W\overline{W})^{m_l+1}\bigr)\biggr)\Biggr] \\
&\qquad\times\frac{1}{N_2+1}\operatorname{tr}\bigl(\boldsymbol{S}_{(\alpha\beta)}\boldsymbol{S}_{(\zeta\eta)\overline{(\lambda\mu)}}(W\overline{W})^{N_2+1}\bigr)\frac{1}{N_3+1}\operatorname{tr}\bigl(\boldsymbol{S}_{\overline{(\gamma\delta)}}\boldsymbol{S}_{(\phi\psi)\overline{(\sigma\tau)}}(W\overline{W})^{N_3+1}\bigr),
\end{aligned}
$$

where $\alpha_1, \beta_1, \ldots, \alpha_{l-1}, \beta_{l-1}, \lambda, \mu, \phi, \psi$ are free indices from 1 *to* g. (We use the convention that the quantity in square brackets equals $\delta_{\lambda\phi}\delta_{\mu\psi}$ if $N_1 = 0$.)

Proof. We resort to the curvature formula again:

$$\widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}} = -\frac{\partial^2\widetilde{\omega}_{(\zeta\eta)\overline{(\sigma\tau)}}}{\partial W_{\alpha\beta}\partial\overline{W_{\gamma\delta}}} + \sum_{\lambda\le\mu}\sum_{\phi\le\psi}\widetilde{\omega}^{\overline{(\lambda\mu)}(\phi\psi)}\frac{\partial\widetilde{\omega}_{(\zeta\eta)\overline{(\lambda\mu)}}}{\partial W_{\alpha\beta}}\frac{\partial\widetilde{\omega}_{(\phi\psi)\overline{(\sigma\tau)}}}{\partial\overline{W_{\gamma\delta}}}.$$

Also from (5-2), we have

$$\begin{aligned}
\frac{\partial^2\widetilde{\omega}_{(\zeta\eta)\overline{(\sigma\tau)}}}{\partial X_{\alpha\beta}\partial\overline{X_{\gamma\delta}}} &= \sum_{k\ge 0}\frac{1}{k+2}\operatorname{tr}\big(\boldsymbol{S}_{(\alpha\beta)\overline{(\gamma\delta)}}\boldsymbol{S}_{(\zeta\eta)\overline{(\sigma\tau)}}(W\overline{W})^{k+2}\big),\\
\frac{\partial\widetilde{\omega}_{(\zeta\eta)\overline{(\lambda\mu)}}}{\partial X_{\alpha\beta}} &= \sum_{k\ge 1}\frac{1}{k+1}\operatorname{tr}\big(\boldsymbol{S}_{(\alpha\beta)}\boldsymbol{S}_{(\zeta\eta)\overline{(\lambda\mu)}}(W\overline{W})^{k+1}\big),\\
\frac{\partial\widetilde{\omega}_{(\phi\psi)\overline{(\sigma\tau)}}}{\partial\overline{X_{\gamma\delta}}} &= \sum_{k\ge 1}\frac{1}{k+1}\operatorname{tr}\big(\boldsymbol{S}_{\overline{(\gamma\delta)}}\boldsymbol{S}_{(\phi\psi)\overline{(\sigma\tau)}}(W\overline{W})^{k+1}\big).
\end{aligned}$$

From this and Lemma 5.2, the result follows. □

Based on Theorem 5.3, the holomorphic sectional curvature $\mathrm{H}(V)$ of $\widetilde{\omega}_s$ along the direction $V=\sum_{\alpha\le\beta}V_{\alpha\beta}\partial/\partial W_{\alpha\beta}$ at the zero matrix of $\mathscr{B}_g$ can be easily gotten:

$$\tag{5-4}\begin{aligned}
\mathrm{H}(V) &= \frac{\sum_{\alpha\le\beta}\sum_{\gamma\le\delta}\sum_{\zeta\le\eta}\sum_{\sigma\le\tau}V_{\alpha\beta}\overline{V_{\gamma\delta}}V_{\zeta\eta}\overline{V_{\sigma\tau}}\widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}}(0)}{\big(\sum_{\alpha\le\beta}\sum_{\gamma\le\delta}\widetilde{\omega}_{(\alpha\beta)\overline{(\gamma\delta)}}(0)V_{\alpha\beta}\overline{V_{\gamma\delta}}\big)^2}\\
&= \frac{\sum_{\alpha,\beta}\sum_{\gamma,\delta}\sum_{\zeta,\eta}\sum_{\sigma,\tau}V_{\alpha\beta}\overline{V_{\gamma\delta}}V_{\zeta\eta}\overline{V_{\sigma\tau}}\widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}}(0)}{\big(\sum_{\alpha,\beta}\sum_{\gamma,\delta}V_{\alpha\beta}\overline{V_{\gamma\delta}}\Delta_{(\alpha\beta)\overline{(\gamma\delta)}}\big)^2}\\
&= -2\frac{\sum_{\alpha,\beta,\gamma,\delta}V_{\alpha\beta}\overline{V_{\beta\gamma}}V_{\gamma\delta}\overline{V_{\delta\alpha}}}{\big(\sum_{\alpha,\beta}V_{\alpha\beta}\overline{V_{\alpha\beta}}\big)^2}.
\end{aligned}$$

In the second equality, $V_{\alpha\beta}$ has a symmetric extension to the whole matrix, and there $(\alpha\beta)$ in $\widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}}$ stands for the derivative with respect to $W_{\alpha\beta}$, not to both $W_{\alpha\beta}$ and $W_{\beta\alpha}$.

By the mean value inequality,

$$\frac{1}{g}\le\frac{\sum_{\alpha,\beta,\gamma,\delta}V_{\alpha\beta}\overline{V_{\beta\gamma}}V_{\gamma\delta}\overline{V_{\delta\alpha}}}{\big(\sum_{\alpha,\beta}V_{\alpha\beta}\overline{V_{\alpha\beta}}\big)^2}\le 1.$$

Thus

$$-2\le\mathrm{H}(V)\le-\frac{2}{g}.$$

Yin [2010, Corollaries 1.1 and 1.2 of Chapter 4] has reproved the classical fact that $\mathrm{H}(V)=-2$ if and only if V is a symmetric matrix of rank 1.

Proposition 5.4. *If* $\mathrm{H}(V) = -2/g$, *then* $V = kUU^T$ *with* $k > 0$ *and* $U \in \mathrm{M}(g, \mathbb{C})$ *unitary.*

Proof. $\mathrm{H}(V) = -2/g$ forces the following inequalities to become equalities:

$$\begin{aligned}\frac{\sum_{\alpha,\beta,\gamma,\delta} V_{\alpha\beta}\overline{V_{\beta\gamma}}V_{\gamma\delta}\overline{V_{\delta\alpha}}}{\bigl(\sum_{\alpha,\beta} V_{\alpha\beta}\overline{V_{\alpha\beta}}\bigr)^2} &= \frac{\sum_{\alpha,\gamma}\bigl|\sum_\beta V_{\alpha\beta}\overline{V_{\gamma\beta}}\bigr|^2}{\bigl(\sum_{\alpha,\beta}|V_{\alpha\beta}|^2\bigr)^2}\\ &= \frac{\sum_{\alpha=\gamma}\bigl|\sum_\beta V_{\alpha\beta}\overline{V_{\gamma\beta}}\bigr|^2+\sum_{\alpha\neq\gamma}\bigl|\sum_\beta V_{\alpha\beta}\overline{V_{\gamma\beta}}\bigr|^2}{\bigl(\sum_{\alpha,\beta}|V_{\alpha\beta}|^2\bigr)^2}\\ &\geq \frac{\sum_\alpha\bigl|\sum_\beta |V_{\alpha\beta}|^2\bigr|^2}{\bigl(\sum_{\alpha,\beta}|V_{\alpha\beta}|^2\bigr)^2} \geq \frac{(1/g)\bigl(\sum_{\alpha,\beta}|V_{\alpha\beta}|^2\bigr)^2}{\bigl(\sum_{\alpha,\beta}|V_{\alpha\beta}|^2\bigr)^2}.\end{aligned}$$

This is equivalent to $\sum_\beta V_{\alpha\beta}\overline{V_{\gamma\beta}} = 0$ and $\sum_\beta |V_{\alpha\beta}|^2 = \sum_\beta |V_{\gamma\beta}|^2$ for any $\alpha \neq \gamma$. Up to a constant multiple that is a real number bigger than zero, $V_{\alpha\beta}$ is symmetric and unitary. Here is a result from [Mok 1989, p. 70]: If $V \in \mathrm{M}(g, \mathbb{C})$ is complex symmetric, V can be written as

$$U^T \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_g \end{pmatrix} U,$$

with U unitary and $\lambda_i \geq 0$. Also since V is unitary,

$$U^T \begin{pmatrix} \lambda_1^2 & & & \\ & \lambda_2^2 & & \\ & & \ddots & \\ & & & \lambda_g^2 \end{pmatrix} \overline{U} = \mathbb{1}_g,$$

which is equivalent to $\lambda_i^2 = 1$, so $\lambda_i = 1$. Thus $V = kUU^T$ with $k > 0$ and U unitary. □

Yin [2010, Lemma 1.2 and Remark 1.6 of Chapter 4] showed that $\widetilde{\mathrm{R}}$ is very strongly seminegative in the sense of Siu, but not very strongly negative.

Remark 5.5. In the literature one sees $\mathrm{H}(V) \in [-1, -1/g]$, since the Bergmann kernel used is $\rho = -2\log\det(\mathbb{1}_g - \overline{W^T}W)$.

The second fundamental form and the totally geodesic submanifold. Recall that

$$\mathcal{T}or_g \xrightarrow{\mathcal{J}^{\mathrm{tor}}} \mathcal{H}_g \longrightarrow \mathcal{B}_g.$$

The period map $\mathcal{J}^{\mathrm{tor}}$ is given by, from (2-7),

$$\mathcal{J}^{\mathrm{tor}}(t) = (\mathrm{i}\mathbb{1}_g - \mathrm{i}A(t)^T)(\mathbb{1}_g + A(t)^T)^{-1} = \mathrm{i}\frac{\mathbb{1}_g - A(t)^T}{\mathbb{1}_g + A(t)^T},$$

where we normalize the target point to $\mathrm{i}\mathbb{1}_g$. The transformation from $\mathcal{H}_g$ to $\mathcal{B}_g$ is given by

$$W = \frac{\mathbb{1}_g + \mathrm{i}Z}{\mathbb{1}_g - \mathrm{i}Z}.$$

Hence $W = A(t)^T = A(t)$, where $A(t)$ is symmetric.

Together with Theorem 5.3, we have

$$\begin{aligned}
(5\text{-}5)\quad \widetilde{\mathrm{R}}_{i\bar{j}k\bar{l}} &= \sum_{\alpha\le\beta}\sum_{\gamma\le\delta}\sum_{\zeta\le\eta}\sum_{\sigma\le\tau} \widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}} \frac{\partial W_{\alpha\beta}}{\partial t_i}\overline{\frac{\partial W_{\gamma\delta}}{\partial t_j}}\frac{\partial W_{\zeta\eta}}{\partial t_k}\overline{\frac{\partial W_{\sigma\tau}}{\partial t_l}} \\
&= \sum_{\alpha,\beta}\sum_{\gamma,\delta}\sum_{\zeta,\eta}\sum_{\sigma,\tau} \widetilde{\mathrm{R}}_{(\alpha\beta)\overline{(\gamma\delta)}(\zeta\eta)\overline{(\sigma\tau)}} \frac{\partial W_{\alpha\beta}}{\partial t_i}\overline{\frac{\partial W_{\gamma\delta}}{\partial t_j}}\frac{\partial W_{\zeta\eta}}{\partial t_k}\overline{\frac{\partial W_{\sigma\tau}}{\partial t_l}} \\
&= -\sum_{N\ge 0}\frac{1}{N+2}\operatorname{tr}\big(\boldsymbol{S}'_{i\bar{j}}\boldsymbol{S}'_{k\bar{l}}(W\overline{W})^{N+2}\big) \\
&\quad + \sum_{N\ge 0}\sum_{\substack{N_1\ge 0, N_2>0, N_3>0\\ \sum_{i=1}^3 N_i = N+1}}\Bigg[\sum_{\substack{m_i>0, 1\le i\le l\\ m_1+\cdots+m_l=N_1}}(-1)^l \\
&\qquad\qquad \times\left(\frac{1}{m_1+1}\operatorname{tr}\big(\boldsymbol{S}_{(\lambda\mu)\overline{(\alpha_1\beta_1)}}(W\overline{W})^{m_1+1}\big)\right) \\
&\qquad\qquad \times\left(\frac{1}{m_2+1}\operatorname{tr}\big(\boldsymbol{S}_{(\alpha_1\beta_1)\overline{(\alpha_2\beta_2)}}(W\overline{W})^{m_2+1}\big)\right)\cdots \\
&\qquad\qquad \times\left(\frac{1}{m_l+1}\operatorname{tr}\big(\boldsymbol{S}_{(\alpha_{l-1}\beta_{l-1})\overline{(\phi\psi)}}(W\overline{W})^{m_l+1}\big)\right)\Bigg] \\
&\qquad\qquad \times\left[\frac{1}{N_2+1}\operatorname{tr}\big(\boldsymbol{S}'_i\boldsymbol{S}'_{k\overline{(\lambda\mu)}}(W\overline{W})^{N_2+1}\big)\right] \\
&\qquad\qquad \times\left[\frac{1}{N_3+1}\operatorname{tr}\big(\boldsymbol{S}'_{\bar{j}}\boldsymbol{S}'_{(\phi\psi)\bar{l}}(W\overline{W})^{N_3+1}\big)\right].
\end{aligned}$$

In the second equality, we also need a symmetric extension of the indices, as in (5-4). In the third equality, $\boldsymbol{S}_i$ means taking the derivative along $\partial/\partial t_i$, since $W = A(t)$ can be seen as a matrix with variable t; and $\partial/\partial t_i$ and $\partial/\partial t_k$ still run through all W's but can't operate simultaneously on a single W, according to Definition 4.1.

Since the calculation is a little bit complicated, we will present a more detailed formula in the Appendix. Also, the second fundamental form of $\mathcal{J}^{\text{tor}}$, restricted to a nonhyperelliptic locus, can be computed out. However, it is difficult to understand the vanishing of the second fundamental form entirely. Partial results are given as follows.

Proposition 5.6. *The second fundamental form of the nonhyperelliptic locus at the central point is*

$$\Sigma_{i\bar{j}k\bar{l}}(0) = \operatorname{tr}(A_{ik}\overline{A_{jl}}).$$

Proof. From (5-5), we easily have $\widetilde{\mathrm{R}}_{i\bar{j}k\bar{l}}(0) = -\frac{1}{2}\operatorname{tr}\big(\boldsymbol{S}'_{i\bar{j}}\boldsymbol{S}'_{k\bar{l}}\big(A_1(t)\overline{A_1(t)}A_1(t)\overline{A_1(t)}\big)\big)$. Also, Theorem 4.4 tells us

$$\begin{aligned}\mathrm{R}_{i\bar{j}k\bar{l}}(0) = &-\operatorname{tr}\big(\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}(A_2(t)\overline{A_2(t)})\big) - \tfrac{1}{2}\operatorname{tr}\big(\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}(A_1(t)\overline{A_1(t)}A_1(t)\overline{A_1(t)})\big)\\ &+\big[\operatorname{tr}\big(\boldsymbol{S}_i\boldsymbol{S}_{k\bar{p}}(A_2(t)\overline{A_1(t)})\big)\big]\big[\operatorname{tr}\big(\boldsymbol{S}_{\bar{j}}\boldsymbol{S}_{p\bar{l}}(A_1(t)\overline{A_2(t)})\big)\big].\end{aligned}$$

Hence $\Sigma_{i\bar{j}k\bar{l}}(0) = \operatorname{tr}(A_{ik}\overline{A_{jl}}) - \operatorname{tr}(A_{ik}\overline{A_p})\operatorname{tr}(A_p\overline{A_{jl}})$. By the use of (4-5), this proposition follows easily. □

Corollary 5.7. *Holomorphic sectional curvature along a totally geodesic submanifold M in a nonhyperelliptic locus of $\mathcal{T}_g$ is bounded from below.*

This shows indicates that a totally geodesic submanifold can't be arbitrarily negatively curved.

Proof. Proposition 5.6 tells us that $A_{ij} = 0$ on the totally geodesic manifold M. From (4-7) and (4-8), the holomorphic sectional curvature $\mathrm{H}(v)$ becomes

$$\mathrm{H}(v) = -2\frac{\operatorname{tr}\big((a_iA_i)\overline{(a_jA_j)}(a_kA_k)\overline{(a_lA_l)}\big)}{\big(\operatorname{tr}((a_iA_i)\overline{(a_jA_j)})\big)^2}.$$

By (4-9),

$$\mathrm{H}(v) \geq -2. \qquad \square$$

Proposition 5.8. *The totally geodesic manifold M in the nonhyperelliptic locus of $\mathcal{T}_g$ must be locally symmetric.*

Proof. We just need to check that $\nabla\mathrm{R} = 0$ on M. We use the well known formulas

$$\begin{aligned}\nabla_p\mathrm{R}_{i\bar{j}k\bar{l}} &= \frac{\partial}{\partial t_p}\mathrm{R}_{i\bar{j}k\bar{l}} - \Gamma^q_{pi}\mathrm{R}_{q\bar{j}k\bar{l}} - \Gamma^q_{pk}\mathrm{R}_{i\bar{j}q\bar{l}},\\ \nabla_{\bar{p}}\mathrm{R}_{i\bar{j}k\bar{l}} &= \frac{\partial}{\partial \bar{t}_p}\mathrm{R}_{i\bar{j}k\bar{l}} - \overline{\Gamma^q_{pj}}\mathrm{R}_{i\bar{q}k\bar{l}} - \overline{\Gamma^q_{pl}}\mathrm{R}_{i\bar{j}k\bar{q}}.\end{aligned}$$

The normal coordinate gives us $\Gamma^k_{ij}(0) = 0$, and thus both $\nabla_p\mathrm{R}_{i\bar{j}k\bar{l}}(0)$ and $\nabla_{\bar{p}}\mathrm{R}_{i\bar{j}kl}(0)$ concern the first derivative of $\mathrm{R}_{i\bar{j}k\bar{l}}$. By Proposition 5.6, $A_{ij} = 0$, i.e., $A_2(t) = 0$, follows from the fact that the manifold M is totally geodesic. Also, Theorem 4.4

implies that

$$
\begin{aligned}
\mathrm{R}^{(1)}_{i\bar{j}k\bar{l}} = &- \operatorname{tr}\bigl(\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}\bigl(A_2(t)\overline{A_3(t)} + A_3(t)\overline{A_2(t)}\bigr)\bigr) \\
&- \operatorname{tr}\bigl(\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}\bigl(A_2(t)\overline{A_1(t)}A_1(t)\overline{A_1(t)} + A_1(t)\overline{A_2(t)}A_1(t)\overline{A_1(t)}\bigr)\bigr) \\
&+ \operatorname{tr}\bigl(\boldsymbol{S}_i\boldsymbol{S}_{k\bar{p}}\bigl(\tfrac{1}{2}A_1(t)\overline{A_1(t)}A_1(t)\overline{A_1(t)} + A_3(t)\overline{A_1(t)}\bigr)\bigr)\operatorname{tr}\bigl(\boldsymbol{S}_{\bar{j}}\boldsymbol{S}_{p\bar{l}}(A_1(t)\overline{A_2(t)})\bigr) \\
&+ \operatorname{tr}\bigl(\boldsymbol{S}_i\boldsymbol{S}_{k\bar{p}}(A_2(t)\overline{A_1(t)})\bigr)\operatorname{tr}\bigl(\boldsymbol{S}_{\bar{j}}\boldsymbol{S}_{p\bar{l}}\bigl(\tfrac{1}{2}A_1(t)\overline{A_1(t)}A_1(t)\overline{A_1(t)} + A_1(t)\overline{A_3(t)}\bigr)\bigr) \\
&- \operatorname{tr}\bigl(\boldsymbol{S}_{q\bar{p}}\bigl(A_2(t)\overline{A_1(t)} + A_1(t)\overline{A_2(t)}\bigr)\bigr) \\
&\qquad\qquad \operatorname{tr}\bigl(\boldsymbol{S}_i\boldsymbol{S}_{k\bar{q}}(A_2(t)\overline{A_1(t)})\bigr)\operatorname{tr}\bigl(\boldsymbol{S}_{\bar{j}}\boldsymbol{S}_{p\bar{l}}(A_1(t)\overline{A_2(t)})\bigr).
\end{aligned}
$$

Every summand above has an $A_2(t)$ term. So $\mathrm{R}^{(1)}_{i\bar{j}k\bar{l}} = 0$, and thus $\nabla_p \mathrm{R}_{i\bar{j}k\bar{l}}(0) = 0$ and $\nabla_{\bar{p}} \mathrm{R}_{i\bar{j}kl}(0) = 0$. □

Appendix

We give here the full formula for the second fundamental form. Let E^{α}_{β} be defined by $(E^{\alpha}_{\beta})_{\gamma\delta} = \delta_{\alpha\gamma}\delta_{\beta\delta}$ (Kronecker delta). The following example will serve to illustrate the meaning of the symbol $\boldsymbol{P}$ (permutation summation):

$$
\begin{aligned}
&\boldsymbol{P}(A\overline{B}C\overline{D}E\overline{F}G\overline{H}) \\
&\qquad = A\overline{B}C\overline{D}E\overline{F}G\overline{H} + C\overline{B}A\overline{D}E\overline{F}G\overline{H} + E\overline{B}C\overline{D}A\overline{F}G\overline{H} + G\overline{B}C\overline{D}E\overline{F}A\overline{H},
\end{aligned}
$$

where A to H are all complex matrices. The matrices with bars are stationary, and those with bars travel through those without bars.

Replace W by $A(t)$ in (5-5) and rearrange that formula according to degrees in t to get

$$
\begin{aligned}
\widetilde{\mathrm{R}}_{i\bar{j}k\bar{l}} = &- \sum_{N\ge 0}\ \sum_{\substack{m_i>0,\,1\le i\le 2l,\,l\ge 2\\ m_1+\cdots+m_{2l}=N+4}} \frac{1}{l}\operatorname{tr}\bigl(\boldsymbol{S}'_{i\bar{j}}\boldsymbol{S}'_{k\bar{l}}\bigl(A_{m_1}(t)\dots\overline{A_{m_{2l}}(t)}\bigr)\bigr) \\
&+ \sum_{\substack{N_1\ge 0,\,N_2>0,\,N_3>0\\ \sum_{i=1}^3 N_i=N}} \Biggl[\ \sum_{\substack{m_i>0,\,1\le i\le l\\ m_1+\cdots+m_l=N_1}} (-1)^l \\
&\qquad \times \sum_{\substack{m_{1n}>0,\\ \sum_{n=1}^{2s_1} m_{1n}=m_1}} \operatorname{tr}\bigl(\boldsymbol{P}\bigl(E^{\lambda}_{\mu}\overline{A_{m_{11}}(t)}\dots A_{m_{12s_1}}(t)\overline{E^{\alpha_1}_{\beta_1}}\bigr)\bigr) \\
&\qquad \times \sum_{\substack{m_{2n}>0,\\ \sum_{n=1}^{2s_2} m_{2n}=m_2}} \operatorname{tr}\bigl(\boldsymbol{P}\bigl(E^{\alpha_1}_{\beta_1}\overline{A_{m_{21}}(t)}\dots A_{m_{22s_2}}(t)\overline{E^{\alpha_2}_{\beta_2}}\bigr)\bigr)\cdots \\
&\qquad \times \sum_{\substack{m_{ln}>0,\\ \sum_{n=1}^{2s_l} m_{ln}=m_l}} \operatorname{tr}\bigl(\boldsymbol{P}\bigl(E^{\alpha_{l-1}}_{\beta_{l-1}}\overline{A_{m_{l1}}(t)}\dots A_{m_{l2s_l}}(t)\overline{E^{\phi}_{\psi}}\bigr)\bigr)\Biggr]
\end{aligned}
$$

$$\times \Bigg[\sum_{\substack{m_i>0,\\ \sum_{i=1}^{2l+1} m_i=N_2+2}} \operatorname{tr}\Big(\boldsymbol{S}_i'\boldsymbol{S}_k'\big(A_{m_1}(t)\dots A_{m_{2l+1}}(t)\overline{E_\mu^\lambda}\big)\Big)\Bigg]$$
$$\times \Bigg[\sum_{\substack{m_i>0,\\ \sum_{i=1}^{2l+1} m_i=N_3+2}} \operatorname{tr}\Big(\boldsymbol{S}_{\bar{j}}'\boldsymbol{S}_{\bar{l}}'\big(E_\psi^\phi\overline{A_{m_1}(t)}\dots \overline{A_{m_{2l+1}}(t)}\big)\Big)\Bigg].$$

By convention, the term in the first square brackets has the value $\delta_{\lambda\phi}\delta_{\mu\psi}$ if $N_1=0$.

Theorem A.9. *The second fundamental form $\Sigma_{i\bar{j}k\bar{l}}$ of the nonhyperelliptic locus is*

$$\begin{aligned}
\Sigma_{i\bar{j}k\bar{l}}
&= \sum_{N\geq 0}\frac{1}{l}\Bigg(\sum_{\substack{m_i>0,1\leq i\leq 2l\\ \sum_{i=1}^{2l} m_i=N+4}} \operatorname{tr}\Big(\boldsymbol{S}_{i\bar{j}}\boldsymbol{S}_{k\bar{l}}\big(A_{m_1}(t)\overline{A_{m_2}(t)}\dots\overline{A_{m_{2l}}(t)}\big)\Big)\\
&\qquad\qquad - \sum_{\substack{m_i>0,l\geq 2\\ \sum_{i=1}^{2l} m_i=N+4}} \operatorname{tr}\Big(\boldsymbol{S}_{i\bar{j}}'\boldsymbol{S}_{k\bar{l}}'\big(A_{m_1}(t)\dots\overline{A_{m_{2l}}(t)}\big)\Big)\Bigg)\\
&\quad + \sum_{N\geq 0}\sum_{\substack{N_1\geq 0,N_2>0,N_3>0\\ \sum_{i=1}^{3} N_i=N}}\Bigg[\sum_{\substack{m_i>0,1\leq i\leq l\\ m_1+\dots+m_l=N_1}}(-1)^l\\
&\qquad \times \sum_{\substack{m_{1n}>0,\\ \sum_{n=1}^{2s_1} m_{1n}=m_1}} \operatorname{tr}\Big(\boldsymbol{P}\big(E_\mu^\lambda\overline{A_{m_{11}}(t)}\dots A_{m_{12s_1}}(t)\overline{E_{\beta_1}^{\alpha_1}}\big)\Big)\\
&\qquad \times \sum_{\substack{m_{2n}>0\\ \sum_{n=1}^{2s_2} m_{2n}=m_2}} \operatorname{tr}\Big(\boldsymbol{P}\big(E_{\beta_1}^{\alpha_1}\overline{A_{m_{21}}(t)}\dots A_{m_{22s_2}}(t)\overline{E_{\beta_2}^{\alpha_2}}\big)\Big)\cdots\\
&\qquad \times \sum_{\substack{m_{ln}>0\\ \sum_{n=1}^{2s_l} m_{ln}=m_l}} \operatorname{tr}\Big(\boldsymbol{P}\big(E_{\beta_{l-1}}^{\alpha_{l-1}}\overline{A_{m_{l1}}(t)}\dots A_{m_{l2s_l}}(t)\overline{E_\psi^\phi}\big)\Big)\Bigg]\\
&\qquad \times \Bigg[\sum_{\substack{m_i>0\\ \sum_{i=1}^{2l+1} m_i=N_2+2}} \operatorname{tr}\Big(\boldsymbol{S}_i'\boldsymbol{S}_k'\big(A_{m_1}(t)\dots A_{m_{2l+1}}(t)\overline{E_\mu^\lambda}\big)\Big)\Bigg]\\
&\qquad \times \Bigg[\sum_{\substack{m_i>0,\\ \sum_{i=1}^{2l+1} m_i=N_3+2}} \operatorname{tr}\Big(\boldsymbol{S}_{\bar{j}}'\boldsymbol{S}_{\bar{l}}'\big(E_\psi^\phi\overline{A_{m_1}(t)}\dots \overline{A_{m_{2l+1}}(t)}\big)\Big)\Bigg]\\
&\quad - \sum_{N\geq 0}\sum_{\substack{N_i\geq 0,1\leq i\leq 3\\ \sum_{i=1}^{3} N_i=N}}\Bigg[\sum_{\substack{m_i>0,1\leq i\leq l\\ m_1+\dots+m_l=N_1}}(-1)^l\prod_{i=1}^{l}\frac{1}{s_i}
\end{aligned}$$

$$
\begin{aligned}
&\times \sum_{\substack{m_{1n}>0,1\le n\le 2s_1\\ \sum_{n=1}^{2s_1} m_{1n}=m_1+2}} \operatorname{tr}\Big(\boldsymbol{S}_{q\bar{i}_1}\big(A_{m_{11}}(t)\dots\overline{A_{m_{12s_1}}(t)}\big)\Big)\\
&\times \sum_{\substack{m_{2n}>0,1\le n\le 2s_2\\ \sum_{n=1}^{2s_2} m_{2n}=m_2+2}} \operatorname{tr}\Big(\boldsymbol{S}_{i_1\bar{i}_2}\big(A_{m_{21}}(t)\dots\overline{A_{m_{22s_2}}(t)}\big)\Big)\cdots\\
&\times\cdots \sum_{\substack{m_{ln}>0,1\le n\le 2s_l\\ \sum_{n=1}^{2s_l} m_{ln}=m_l+2}} \operatorname{tr}\Big(\boldsymbol{S}_{i_{l-1}\bar{p}}\big(A_{m_{l1}}(t)\dots\overline{A_{m_{l2s_l}}(t)}\big)\Big)\Bigg]\\
\times\Bigg[&\sum_{\substack{m_i>0,1\le i\le 2l\\ \sum_{i=1}^{2l} m_i=N_2+3}} \frac{1}{l}\operatorname{tr}\Big(\boldsymbol{S}_i\boldsymbol{S}_{k\bar{q}}\big(A_{m_1}(t)\dots\overline{A_{m_{2l}}(t)}\big)\Big)\Bigg]\\
\times\Bigg[&\sum_{\substack{m_i>0,1\le i\le 2l\\ \sum_{i=1}^{2l} m_i=N_3+3}} \frac{1}{l}\operatorname{tr}\Big(\boldsymbol{S}_{\bar{j}}\boldsymbol{S}_{p\bar{l}}\big(A_{m_1}(t)\dots\overline{A_{m_{2l}}(t)}\big)\Big)\Bigg].
\end{aligned}
$$

Proof. This follows from Theorem 4.4, the formula for $\widetilde{\mathrm{R}}_{i\bar{j}k\bar{l}}$ above and (5-1). □

Acknowledgements

The authors thank their advisor, Professor Kefeng Liu, for his constant help and encouragement. They are also grateful to Dr. Fangliang Yin, Professor Fangyang Zheng and Dr. Shengmao Zhu for their useful discussions and comments. The referees' useful suggestions and the editors' excellent work are also appreciated.

References

[Ahlfors 1966] L. V. Ahlfors, *Lectures on quasiconformal mappings*, Van Nostrand, Princeton, 1966. Reprinted Brooks/Cole 1987 and AMS 2006. MR 2009d:30001 Zbl 1103.30001

[Bojarski 1955] B. V. Bojarski, "Homeomorphic solutions of Beltrami systems", *Dokl. Akad. Nauk SSSR* (*N.S.*) **102** (1955), 661–664. In Russian. MR 17,157a

[Colombo and Frediani 2010] E. Colombo and P. Frediani, "Siegel metric and curvature of the moduli space of curves", *Trans. Amer. Math. Soc.* **362**:3 (2010), 1231–1246. MR 2010h:14046 Zbl 1196.14025

[Farb 2006] B. Farb (editor), *Problems on mapping class groups and related topics*, Proceedings of Symposia in Pure Mathematics **74**, Amer. Math. Soc., Providence, RI, 2006. MR 2007e:57001 Zbl 1098.58001

[Farkas and Kra 1992] H. M. Farkas and I. Kra, *Riemann surfaces*, 2nd ed., Graduate Texts in Mathematics **71**, Springer, New York, 1992. MR 93a:30047 Zbl 0764.30001

[Hain 1999] R. Hain, "Locally symmetric families of curves and Jacobians", pp. 91–108 in *Moduli of curves and abelian varieties*, edited by C. Faber and E. Looijenga, Aspects Math. **E33**, Vieweg, Braunschweig, 1999. MR 2000i:14040 Zbl 0966.14018

[Liu et al. 2012a] K. F. Liu, S. Rao, and X. K. Yang, “Hodge theory and deformations of Kähler manifolds”, preprint, 2012. arXiv 1207.1182

[Liu et al. 2012b] K. F. Liu, Q. T. Zhao, and S. Rao, “New proofs of the Torelli theorems for Riemann surfaces”, preprint, 2012. arXiv 1207.5697

[Mayer 1969] A. Mayer, “Rauch’s variational formula and the heat equation”, *Math. Ann.* **181** (1969), 53–59. MR 40 #1588 Zbl 0159.22403

[Mok 1989] N. Mok, *Metric rigidity theorems on Hermitian locally symmetric manifolds*, Series in Pure Mathematics **6**, World Scientific, Teaneck, NJ, 1989. MR 92d:32046 Zbl 0912.32026

[Moonen 1995] B. J. J. Moonen, *Special points and linearity properties of Shimura varieties*, thesis, University of Utrecht, 1995.

[Moonen and Oort 2013] B. J. J. Moonen and F. Oort, “The Torelli locus and special subvarieties”, pp. 549–594 in *Handbook of moduli*, vol. 2, edited by G. Farkas and I. Morrison, Advanced Lectures in Mathematics **25**, International Press, Somerville, MA, 2013.

[Morrey 1938] C. B. Morrey, Jr., “On the solutions of quasi-linear elliptic partial differential equations”, *Trans. Amer. Math. Soc.* **43**:1 (1938), 126–166. MR 1501936 Zbl 0018.40501

[Oort 1997] F. Oort, “Canonical liftings and dense sets of CM-points”, pp. 228–234 in *Arithmetic geometry* (Cortona, 1994), edited by F. Catanese, Sympos. Math. **37**, Cambridge University Press, Cambridge, 1997. MR 98g:14056 Zbl 0911.14018

[Oort 2003] F. Oort, “Special points in Shimura varieties: an introduction”, lecture notes, 2003, http://www.staff.science.uu.nl/~oort0109/A-IC.SpecialPoints.ps.

[Rauch 1959] H. E. Rauch, “Weierstrass points, branch points, and moduli of Riemann surfaces”, *Comm. Pure Appl. Math.* **12** (1959), 543–560. MR 22 #1666 Zbl 0091.07301

[Toledo 1987] D. Toledo, “Nonexistence of certain closed complex geodesics in the moduli space of curves”, *Pacific J. Math.* **129**:1 (1987), 187–192. MR 88j:32026 Zbl 0648.14018

[Yin 2010] F. L. Yin, *Geometry of period map of Riemann surface*, thesis, Zhejiang University, 2010.

Received August 11, 2012.

QUANTING ZHAO
SCHOOL OF MATHEMATICS AND STATISTICS
CENTRAL CHINA NORMAL UNIVERSITY
WUHAN 430079
CHINA
zhaoquanting@126.com

SHENG RAO
SCHOOL OF MATHEMATICS AND STATISTICS
WUHAN UNIVERSITY
WUHAN 430072
CHINA
likeanyone@zju.edu.cn

Past affiliation of both authors:

CENTER OF MATHEMATICAL SCIENCES
ZHEJIANG UNIVERSITY
HANGZHOU 310027
CHINA

Guidelines for Authors

Authors may submit manuscripts at msp.berkeley.edu/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 266 No. 1 November 2013
